# MINIMAX-OPTIMAL MULTICLASS CLASSIFICATION WITH CYCLIC OBLIQUE FORESTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Oblique splits can represent classification boundaries that are poorly aligned with the input coordinates. A key challenge is to combine this geometric flexibility with sharp statistical guarantees that exploit smooth class probabilities. We introduce the Cyclic Oblique Forest, a purely randomized approach to multiclass classification that combines cyclic oblique splitting with regularized local polynomial estimation. Under uniform design, we establish minimax rates for expected multiclass excess risk over Hölder classes of conditional class probabilities for every fixed finite smoothness order, without additional logarithmic factors. These rates are attained with midpoint cuts along a basis of splitting directions, with depth and polynomial degree chosen for the specified smoothness. We further derive a sharper excess-risk bound under an additional multiclass margin condition with a positive exponent. Real data analysis on classic UCI datasets shows that the proposed method achieves lower mean test error than existing random forest baselines.

## 1 INTRODUCTION

Random forests are widely used nonparametric methods for classification and regression with strong empirical performance across a broad range of applications (Breiman, 2001; Gislason et al., 2006; Qi, 2012; Khaidem et al., 2016). Standard implementations combine predictions from randomized decision trees built by recursively splitting on individual features to form axis-aligned partitions. This geometric restriction can require many splits to approximate even a linear decision boundary that is not aligned with the coordinate axes (Murthy et al., 1994). Oblique forests relax this restriction by allowing splits on linear combinations of features. Representative approaches obtain splitting directions through learned projections (Menze et al., 2011; Rainforth & Wood, 2015), random rotations (Blaser & Fryzlewicz, 2016), or sparse random projections (Tomita et al., 2020). These constructions allow individual splits to align with boundaries that would otherwise require a finer axis-aligned partition. As an example, consider $\mathbf{X} \sim \mathrm{Unif}([0,1]^2)$ with conditional class probabilities

$$\eta_1(\mathbf{x}) = 0.5 + 0.4(x_1 + x_2 - 1), \quad \eta_2(\mathbf{x}) = 1 - \eta_1(\mathbf{x}), \quad \mathbf{x} = (x_1, x_2)^\top. \tag{1}$$

The Bayes decision boundary is $x_1 + x_2 = 1$. Figure 1 compares midpoint partitions whose leaves predict their population majority class. A single oblique midpoint split along this boundary yields zero disagreement probability with the Bayes classifier, whereas alternating coordinate midpoint splits yield a disagreement probability of $3.125\%$ at depth eight. Despite these geometric advantages, existing convergence guarantees for oblique forests remain limited in scope. This motivates a broader statistical question: can we construct oblique forest classifiers that attain minimax-optimal rates for expected multiclass excess risk over Hölder classes of conditional class probabilities with arbitrary finite smoothness?

Let $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be a training sample of independent and identically distributed observations from a joint distribution $P$ of $(\mathbf{X}, Y)$, where $\mathbf{X} \in [0,1]^d$ is the feature vector and $Y \in \{1, \ldots, \tau\}$ is the class label, for an integer $\tau \geq 2$. We focus on multiclass classification with conditional class probabilities $\eta_c(\mathbf{x}) = \mathbb{P}(Y = c \mid \mathbf{X} = \mathbf{x})$, $c = 1, \ldots, \tau$. We evaluate a random forest classifier $\widehat{g}$ trained on $\mathcal{D}_n$ by its expected excess misclassification risk $\mathbb{E}[R(\widehat{g}) - R(g^*)]$, where $R(g) = \mathbb{P}(g(\mathbf{X}) \neq Y)$ is the risk on an independent test observation, and $g^*(\mathbf{x}) \in \arg\max_{1 \leq c \leq \tau} \eta_c(\mathbf{x})$ is a Bayes classifier. The expectation is taken over the training sample and the randomization used to construct the forest.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

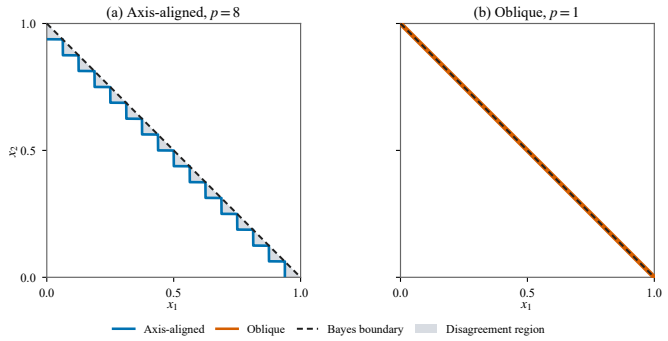


Figure 1: **Geometric cost of axis alignment.** (a) Axis-aligned leaves crossed by the diagonal Bayes boundary contain points with different Bayes-optimal labels. Assigning a single class to each leaf produces the shaded disagreement regions. (b) An oblique split separates the two Bayes decision regions exactly. Here $p$ denotes tree depth, with the root at depth zero.

Our theoretical focus is on purely randomized forests with partitions generated independently of the training sample and leaf predictions fitted using the observations. This separation allows us to analyze how partition geometry and leaf estimation jointly determine risk. Early analyses established consistency for purely randomized forests (Biau et al., 2008) and convergence rates for centered and uniformly random partitions (Biau, 2012; Genuer, 2012). Subsequent work examined the approximation benefits of aggregation (Arlot & Genuer, 2014), connections with kernel estimators (Scornet, 2016; Iakovidis & Arcozzi, 2023), and sharper rates for centered forests (Klusowski, 2021). These studies highlight the role of coordinate split allocation. More recently, Bettinger et al. (2026) relate elongated cells produced by independent coordinate selection to super-logarithmic corrections in pointwise error bounds. Ehrenfest centered forests (Neumeyer et al., 2025) control this allocation and attain optimal regression rates over Hölder classes with smoothness order $s \in (0, 1]$. For classification, Gao & Zhou (2020); Gao et al. (2022) establish convergence bounds for simplified axis-aligned forests in binary and multiclass settings under a Lipschitz assumption on the conditional class probabilities. Mondrian forests (Lakshminarayanan et al., 2014) provide another framework for consistency and minimax analysis (Mourtada et al., 2017; 2020). Their regression theory establishes optimal rates for trees when $0 < s \leq 1$ and for sufficiently large forests when $1 < s \leq 2$, with interior restrictions for guarantees covering the latter range. Related developments include online aggregation (Mourtada et al., 2021) and extensions to general loss functions (Zhan et al., 2024).

Axis-aligned constructions can also exploit higher-order smoothness. Cai et al. (2023) establish regression guarantees for arbitrary finite smoothness, with logarithmic factors in their convergence bounds. Debiased Mondrian forests (Cattaneo et al., 2026) attain the minimax mean squared error rate for arbitrary fixed finite $s > 0$ at interior points and over fixed interior regions. With sample-dependent partitions, local adaptive split balancing forests (Zhang et al., 2024) combine balanced coordinate splitting with polynomial leaf estimation and achieve minimax-order integrated squared regression error in probability for arbitrary finite Hölder smoothness.

For oblique partitions, existing minimax analyses rely primarily on random tessellations. Building on the STIT framework of O'Reilly & Tran (2022), O'Reilly & Tran (2024) establish minimax regression rates and corresponding binary classification bounds for STIT and Poisson hyperplane forests with smoothness up to order two, under suitable design and boundary conditions. Oblique Mondrian forests (O'Reilly, 2024) can attain rates governed by the dimension of a relevant subspace when their splitting directions approximate that subspace sufficiently accurately. Online binary space partitioning forests (Fan et al., 2020) attain optimal regression rates for Lipschitz functions, while uniformly rotated Mondrian kernels (Osborne & O'Reilly, 2025) provide a related construction with a rotation-invariant limiting kernel. Table 1 compares representative guarantees across partition geometries, risk criteria, and smoothness ranges. To the best of our knowledge, existing theory for oblique random forests does not establish minimax rates without additional logarithmic factors for expected multiclass excess risk over Hölder classes of arbitrary fixed finite smoothness, with risk evaluated over the full input domain.

We address this gap with the *Cyclic Oblique Forest*, whose trees cycle through a fixed collection of splitting directions and fit regularized local polynomial class scores within the resulting leaves.

Table 1: **Convergence rates for random forests.**

| Reference | Geometry | Risk | Smoothness | Convergence rate |
|---|---|---|---|---|
| Gao & Zhou (2020) | Axis-aligned | Binary excess risk | Lipschitz | $\mathcal{O}((\log n/n)^{1/(d+2)})$ |
| Neumeyer et al. (2025) | Axis-aligned | Pointwise RMSE | $0 < s \leq 1$ | $\mathcal{O}(n^{-s/(2s+d)})$ |
| Mourtada et al. (2020)† | Axis-aligned | Binary excess risk | $0 < s \leq 2$ | $\mathcal{O}(n^{-s/(2s+d)})$ |
| Cai et al. (2023) | Axis-aligned | RIMSE | $s > 0$ | $\mathcal{O}((\log n/n)^{s/(2s+d)})$ |
| Cattaneo et al. (2026)† | Axis-aligned | Pointwise RMSE | $s > 0$ | $\mathcal{O}(n^{-s/(2s+d)})$ |
| O'Reilly & Tran (2024)† | Oblique tessellation | Binary excess risk | $1 < s \leq 2$ | $\mathcal{O}(n^{-s/(2s+d)})$ |
| O'Reilly (2024)† | Oblique tessellation | RIMSE | $1 < s \leq 2$ | $\mathcal{O}(n^{-s/(2s+r)})$ |
| Fan et al. (2020) | Oblique recursive | RIMSE | Lipschitz | $\mathcal{O}(n^{-1/(d+2)})$ |
| **This work** | **Oblique recursive** | **Multiclass excess risk** | $s > 0$ | $\boldsymbol{\mathcal{O}(n^{-s/(2s+d)})}$ |

Let $f : [0,1]^d \to \mathbb{R}$ be a measurable function and $\widehat{f}$ an estimator of $f$. The root integrated mean squared error (RIMSE) is $(\mathbb{E}\int_{[0,1]^d} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 \, d\mathbf{x})^{1/2}$, with integration taken with respect to Lebesgue measure. The pointwise root mean squared error (RMSE) at $\mathbf{x}$ is $(\mathbb{E}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2)^{1/2}$. Expectations are taken over the training sample and any randomization used to construct the estimator. † denotes guarantees at interior points or over interior regions separated from the domain boundary by a fixed positive distance. $r$ is the aligned subspace dimension.

The forest combines tree predictions by majority vote. The main analytical challenge is to control polynomial estimation error on irregular oblique cells, including leaves with fewer observations than polynomial coefficients or no observations at all. Our contributions are as follows:

- **Geometric control and local polynomial estimation.** We establish deterministic bounds on terminal-cell diameters under cyclic oblique splitting and expected integrated error bounds for regularized polynomial leaf scores under uniform design. The estimation analysis accommodates irregular cells, leaves with few observations, and empty leaves, with constants independent of leaf aspect ratios.
- **Minimax rates for arbitrary finite smoothness.** With centered splits ($\alpha = 1/2$) along $d$ linearly independent directions, we establish matching upper and lower bounds of order $n^{-s/(2s+d)}$ for expected multiclass excess risk over Hölder classes of every fixed finite smoothness order $s > 0$. The polynomial degree and tuning parameters are chosen according to $s$. The guarantee holds over the entire domain $[0,1]^d$ under uniform design and contains no additional logarithmic factor.
- **Margin-dependent bounds and regression extension.** We derive faster excess risk bounds under a multiclass margin condition. The analysis also extends to regression function estimation, yielding the minimax integrated mean squared error rate $n^{-2s/(2s+d)}$; see Appendix A.5.
- **Controlled simulations and predictive evaluation.** We examine the effects of splitting directions and polynomial degree in controlled simulations and compare our method with tuned random forest baselines on three classification tasks and one regression task. The quadratic cyclic oblique forest achieves the lowest mean test error among the evaluated methods on all four tasks.

### 1.1 Notation

Let $\mathbb{N}_0 = \{0, 1, \ldots\}$, and let $\mathbf{1}\{\cdot\}$ denote an indicator function. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer not exceeding $x$. For integers $a \geq 0$ and $b \geq 1$, $a \bmod b$ denotes the remainder when $a$ is divided by $b$. The notation $\mathrm{diag}(a_1, \ldots, a_m)$ denotes the $m \times m$ diagonal matrix with diagonal entries $a_1, \ldots, a_m$. The cardinality of any set $I$ is denoted by $|I|$. For nonnegative sequences $a_n$ and $b_n$, the notation $a_n \lesssim b_n$ denotes $a_n \leq C b_n$ for some constant $C > 0$ and all $n$. The notation $a_n = \mathcal{O}(b_n)$ denotes the same bound for all sufficiently large $n$. The notation $a_n \asymp b_n$ denotes $c b_n \leq a_n \leq C b_n$ for all $n$ and with constants $c, C > 0$.

For $\mathbf{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, define the Euclidean norm as $\|\mathbf{x}\|_2 := (\sum_{j=1}^d x_j^2)^{1/2}$. For $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m \times r}$, the Frobenius and operator norms are defined as $\|\mathbf{A}\|_F := (\sum_{i=1}^m \sum_{j=1}^r a_{ij}^2)^{1/2}$ and $\|\mathbf{A}\|_{\mathrm{op}} := \sup_{\mathbf{y} \in \mathbb{R}^r : \|\mathbf{y}\|_2 = 1} \|\mathbf{A}\mathbf{y}\|_2$, respectively. For a measurable vector-valued function $f : [0,1]^d \to \mathbb{R}^\tau$, we define its $L_2$ norm by $\|f\|_{L_2}^2 = \int_{[0,1]^d} \|f(\mathbf{x})\|_2^2 \, d\mathbf{x}$.

## 2 CYCLIC OBLIQUE FOREST

We introduce the cyclic oblique forest, which combines data-independent oblique partitions with regularized local polynomial class scores and aggregates tree predictions by majority vote. Unless stated otherwise, the feature vector $\mathbf{X}$ has marginal distribution $P_{\mathbf{X}} = \mathrm{Unif}([0,1]^d)$, as in Biau (2012); Scornet et al. (2015); Friedberg et al. (2020); Klusowski (2021); Cai et al. (2023). The dimension $d$ is held fixed throughout the paper.

We describe a single tree that recursively partitions the feature space by cycling through $q$ oblique directions, where $q \geq d$ is an integer. Let $\Theta = (I, T)$ denote the tree randomization. The index set $I \subseteq \{1, \ldots, n\}$ is drawn uniformly without replacement, with size $|I| \in \{1, \ldots, n\}$ satisfying $|I| \asymp n$. The observations indexed by $I$ are used for leaf estimation. The permutation $T$ is drawn uniformly from all permutations of $\{1, \ldots, q\}$ and determines the splitting order. The two draws are independent of each other and of the training sample. In the single-tree construction below, the dependence on $\Theta$ is suppressed in the notation.

The trees share a fixed collection of unit splitting directions $\mathbf{b}_1, \ldots, \mathbf{b}_q \in \mathbb{R}^d$, chosen independently of the training sample. Starting from the root cell $[0,1]^d$ at depth zero, the tree grows to depth $p \in \mathbb{N}_0$. The value $T(k)$ denotes the direction index in position $k$ of the permutation. The directions are visited in the order $\mathbf{b}_{T(1)}, \ldots, \mathbf{b}_{T(q)}$, with the same order repeated every $q$ levels. Let $\alpha \in (0, 1/2]$ denote the split fraction, so that the threshold lies a fraction $\alpha$ of the way from the lower to the upper endpoint of each cell's projection interval. At level $t = 1, \ldots, p$, each cell $K$ at depth $t-1$ is split orthogonally to $\mathbf{b}_{j_t}$ at $m_{j_t}(K) = \alpha \sup_{\mathbf{x} \in K} \mathbf{b}_{j_t}^\top \mathbf{x} + (1-\alpha) \inf_{\mathbf{x} \in K} \mathbf{b}_{j_t}^\top \mathbf{x}$, where $j_t = T(1 + ((t-1) \bmod q))$. The two child cells are

$$K^{\mathrm{L}} = K \cap \{\mathbf{x} : \mathbf{b}_{j_t}^\top \mathbf{x} < m_{j_t}(K)\}, \qquad K^{\mathrm{R}} = K \cap \{\mathbf{x} : \mathbf{b}_{j_t}^\top \mathbf{x} \geq m_{j_t}(K)\}. \tag{2}$$

Repeating this procedure through depth $p$ yields $2^p$ terminal cells, called leaves, which form the partition $\Pi_p = \{A^1, \ldots, A^{2^p}\}$.

Within each leaf $A \in \Pi_p$ of the cyclic partition, we estimate class scores using local polynomials to exploit the smoothness of the conditional class probabilities. For a polynomial degree $\kappa \in \mathbb{N}_0$, let $\mathcal{P}_\kappa(A)$ denote the $D$-dimensional space of polynomials on $A$ with total degree at most $\kappa$, where $D = \binom{d+\kappa}{\kappa}$. For each leaf $A$ with $\mathbb{P}(\mathbf{X} \in A) > 0$, construct a basis $\{\phi_{A,1}, \ldots, \phi_{A,D}\}$ of $\mathcal{P}_\kappa(A)$ satisfying

$$\phi_{A,1} \equiv 1, \qquad \mathbb{E}[\phi_{A,j}(\mathbf{X})\phi_{A,k}(\mathbf{X}) \mid \mathbf{X} \in A] = \mathbf{1}\{j = k\}, \quad j,k = 1, \ldots, D. \tag{3}$$

Define $\boldsymbol{\phi}_A(\mathbf{x}) = (\phi_{A,1}(\mathbf{x}), \ldots, \phi_{A,D}(\mathbf{x}))^\top$ and $\boldsymbol{\eta}(\mathbf{x}) = (\eta_1(\mathbf{x}), \ldots, \eta_\tau(\mathbf{x}))^\top$ for any $\mathbf{x} \in A$. The population coefficient matrix is $\Omega_A = \mathbb{E}[\boldsymbol{\phi}_A(\mathbf{X})\boldsymbol{\eta}(\mathbf{X})^\top \mid \mathbf{X} \in A]$.

For each leaf $A$, let $I_A = \{i \in I : \mathbf{X}_i \in A\}$ and $J = \mathrm{diag}(0, 1, \ldots, 1) \in \mathbb{R}^{D \times D}$. For $|I_A| > 0$ and each class $c = 1, \ldots, \tau$, define

$$\widehat{\boldsymbol{\omega}}_{A,c} = \arg\min_{\boldsymbol{\omega} \in \mathbb{R}^D} \left\{ \frac{1}{|I_A|} \sum_{i \in I_A} \left( \mathbf{1}\{Y_i = c\} - \boldsymbol{\phi}_A(\mathbf{X}_i)^\top \boldsymbol{\omega} \right)^2 + \lambda \boldsymbol{\omega}^\top J \boldsymbol{\omega} \right\}, \tag{4}$$

where $\lambda > 0$ is the ridge parameter. The penalty applies to all coefficients except the intercept. Project the classwise coefficients jointly onto the unit Frobenius ball and define the leaf score vector by

$$\widehat{\Omega}_A = \frac{[\widehat{\boldsymbol{\omega}}_{A,1}, \ldots, \widehat{\boldsymbol{\omega}}_{A,\tau}]}{\max\left\{1, \|[\widehat{\boldsymbol{\omega}}_{A,1}, \ldots, \widehat{\boldsymbol{\omega}}_{A,\tau}]\|_F\right\}}, \qquad \widehat{\boldsymbol{\eta}}(\mathbf{x}) = \widehat{\Omega}_A^\top \boldsymbol{\phi}_A(\mathbf{x}), \qquad \mathbf{x} \in A. \tag{5}$$

The unit Frobenius ball contains the population coefficient matrix $\Omega_A$. Indeed, (3) and Bessel's inequality imply $\|\Omega_A\|_F^2 \leq \mathbb{E}[\|\boldsymbol{\eta}(\mathbf{X})\|_2^2 \mid \mathbf{X} \in A] \leq 1$, where the last inequality follows because $\boldsymbol{\eta}(\mathbf{X})$ is a probability vector. The joint projection rescales all class scores by the same positive factor and therefore leaves the tree's argmax prediction unchanged. For a leaf $A$ with $|I_A| = 0$, use the class proportions of the full tree subsample, $\widehat{p}_{I,c} = |I|^{-1} \sum_{i \in I} \mathbf{1}\{Y_i = c\}$ for $c = 1, \ldots, \tau$, and set $\widehat{\boldsymbol{\eta}}(\mathbf{x}) = (\widehat{p}_{I,1}, \ldots, \widehat{p}_{I,\tau})^\top$ for every $\mathbf{x} \in A$. When $\kappa = 0$, the score vector in each nonempty leaf is exactly its vector of empirical class proportions.

For the cyclic oblique forest, we make the dependence on the tree randomization $\Theta$ explicit and denote the partition and score vector by $\Pi_p(\Theta)$ and $\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta)$, respectively. The tree prediction is $\widehat{g}(\mathbf{x}, \Theta) \in \arg\max_{1 \leq c \leq \tau} \widehat{\eta}_c(\mathbf{x}, \Theta)$, where $\widehat{\eta}_c(\mathbf{x}, \Theta)$ denotes the class-$c$ component of $\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta)$.

For an integer $M \geq 1$, draw independent and identically distributed copies $\Theta_1, \ldots, \Theta_M$ of $\Theta$, independently of the training sample. The forest classifier selects the class with the largest vote proportion:

$$\widehat{g}_M(\mathbf{x}) \in \arg\max_{1 \leq c \leq \tau} \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{\widehat{g}(\mathbf{x}, \Theta_m) = c\}. \tag{6}$$

A fixed deterministic rule resolves ties in both tree and forest predictions. Algorithm 1 in Appendix A.1 summarizes the complete construction.

## 3 Theoretical Analysis

We establish bounds on leaf diameters and polynomial estimation error, then derive excess risk rates and a matching minimax lower bound. We also examine how a multiclass margin condition sharpens the excess risk bound.

### 3.1 Bounds on Leaf Diameters

The cyclic splitting rule contracts the projection widths of terminal cells along the splitting directions. To control local polynomial approximation error, we bound the diameters of the terminal cells. The following condition allows us to transform contraction along the splitting directions into Euclidean diameter bounds.

**Assumption 3.1.** *There exists a finite constant $\Lambda > 0$ such that $\|\mathbf{v}\|_2^2 \leq \Lambda \sum_{j=1}^{q} (\mathbf{b}_j^\top \mathbf{v})^2$ for all $\mathbf{v} \in \mathbb{R}^d$.*

We refer to Assumption 3.1 as the *nondegeneracy condition on the splitting directions*. It is equivalent to $\lambda_{\min}(\sum_{j=1}^{q} \mathbf{b}_j \mathbf{b}_j^\top) \geq \Lambda^{-1}$, where $\lambda_{\min}$ denotes the smallest eigenvalue, and ensures that the splitting directions span $\mathbb{R}^d$. The following examples provide explicit choices of $\Lambda$ for configurations in which the condition holds automatically.

*Example* 3.2. When $q = d$ and $B = [\mathbf{b}_1, \ldots, \mathbf{b}_d]^\top \in \mathbb{R}^{d \times d}$ is invertible, $\sum_{j=1}^{d} (\mathbf{b}_j^\top \mathbf{v})^2 = \|B\mathbf{v}\|_2^2$, and one may take $\Lambda = \|B^{-1}\|_{\mathrm{op}}^2$. In particular, when $B$ is orthogonal, the splitting directions form an orthonormal basis of $\mathbb{R}^d$, and one may take $\Lambda = 1$.

*Example* 3.3. When $q > d$ and the unit directions satisfy $\sum_{j=1}^{q} (\mathbf{b}_j^\top \mathbf{v})^2 = (q/d)\|\mathbf{v}\|_2^2$ for every $\mathbf{v} \in \mathbb{R}^d$, one may take $\Lambda = d/q$. When $q > d$ and the collection contains an orthonormal basis of $\mathbb{R}^d$, the projections onto that basis contribute $\|\mathbf{v}\|_2^2$, while the remaining terms are nonnegative. Thus, one may take $\Lambda = 1$, although this choice need not be optimal.

For a nonempty bounded set $K \subset \mathbb{R}^d$, let $\mathrm{diam}(K) = \sup_{\mathbf{x},\mathbf{y} \in K} \|\mathbf{x} - \mathbf{y}\|_2$ denote its diameter. Along any root-to-leaf path, each complete cycle contracts the projection width along every splitting direction by a factor of at most $1 - \alpha$. A path of depth $p$ contains $\lfloor p/q \rfloor$ complete cycles, yielding the following bound.

**Lemma 3.4.** *Assume that Assumption 3.1 holds. Let $A_p(\mathbf{x})$ denote the unique terminal cell containing $\mathbf{x}$ at depth $p$. For any depth $p \in \mathbb{N}_0$, permutation $T$, and point $\mathbf{x} \in [0,1]^d$,*

$$\mathrm{diam}(A_p(\mathbf{x}))^2 \leq \Lambda q d (1 - \alpha)^{2\lfloor p/q \rfloor}.$$

For fixed splitting directions and $\alpha$, the maximum leaf diameter therefore decays exponentially with depth. This rate will be used to bound the approximation term in the subsequent excess risk analysis.

With centered splits ($\alpha = 1/2$) along the $d$ coordinate directions, our squared-diameter bound has order $\mathcal{O}(2^{-2p/d})$. This matches the order obtained in the geometric analyses of the longest-side constructions of Cai et al. (2023) and Gao & Zhou (2020), which similarly use centered splits and balance coordinate coverage. By contrast, independent uniform coordinate selection in the centered model of Biau (2012) yields expected squared diameter $d(1-3/(4d))^p$. Balanced coordinate coverage avoids elongated leaves from uneven splitting and yields faster squared-diameter decay for $d > 1$.

For centered splits, the lemma gives deterministic squared-diameter bounds of $\mathcal{O}(2^{-2p/d})$ for oblique directions with $q = d$ and $\mathcal{O}(2^{-2p/q})$ for redundant direction sets with $q > d$, simultaneously over

all terminal cells and direction permutations. In comparison, STIT and Poisson hyperplane forests (O'Reilly & Tran, 2024), oblique Mondrian forests (O'Reilly, 2024), and the BSP partition process underlying online BSP forests (Fan et al., 2020) provide pointwise squared-diameter bounds of $\mathcal{O}(2^{-2p/d}\log^2(C/\delta))$ with probability at least $1-\delta$, for $\delta \in (0,1)$ and some fixed constant $C > 1$. These bounds hold for fixed nondegenerate directional distributions when the respective lifetime, intensity, or time budget is of order $2^{p/d}$, giving an expected cell count of $\mathcal{O}(2^p)$. The constants are independent of $p$ and $\delta$. These pointwise bounds do not directly provide simultaneous control over all terminal cells.

### 3.2 Minimax Rates for Multiclass Classification

Smaller leaves reduce polynomial approximation error but contain fewer observations on average, which can increase estimation error. The following lemma controls the expected integrated squared estimation error of the leaf scores conditional on the tree randomization.

**Lemma 3.5.** *Let $\Pi_p = \Pi_p(\Theta)$ and define the population polynomial projection by $\boldsymbol{\eta}_{\Pi_p}(\mathbf{x},\Theta) = \sum_{A\in\Pi_p} \Omega_A^\top \boldsymbol{\phi}_A(\mathbf{x})\mathbf{1}\{\mathbf{x}\in A\}$. For every realization of $\Theta$,*

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{\eta}}(\cdot,\Theta) - \boldsymbol{\eta}_{\Pi_p}(\cdot,\Theta)\right\|_{L_2}^2 \,\middle|\, \Theta\right] \le C_{\mathrm{est}}\left(\frac{2^p}{|I|} + \lambda^2\right),$$

*where $C_{\mathrm{est}} < \infty$ depends only on $(\kappa, d, \tau)$ and the constants implicit in $|I| \asymp n$. Moreover, choosing $\lambda \asymp \sqrt{2^p/n}$ yields*

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{\eta}}(\cdot,\Theta) - \boldsymbol{\eta}_{\Pi_p}(\cdot,\Theta)\right\|_{L_2}^2 \,\middle|\, \Theta\right] = \mathcal{O}(2^p/n).$$

The first term bounds the sampling error across the $2^p$ leaves, while the second controls the bias introduced by ridge regularization. Conditional on $\Theta$, each leaf count is binomial, with success probability equal to the leaf volume under uniform design. Integration weights each leaf's error by its volume, cancelling the inverse-volume factor in the sampling error bound. This yields the stated bound without requiring a minimum leaf probability or a minimum sample count per leaf, and the constant is independent of leaf aspect ratios.

These properties distinguish the roles of geometry and statistical estimation in convergence analysis. Bettinger et al. (2026) relate elongated cells produced by independent coordinate selection in centered and uniform trees to super-logarithmic corrections in pointwise error bounds. Even for the constructions of Gao & Zhou (2020) and Cai et al. (2023), which balance coordinate coverage, the available statistical bounds retain logarithmic factors. For the number of leaves $K$, Gao & Zhou (2020) obtain a leading term $\mathcal{O}(\sqrt{K\log n/n})$ in expected binary excess risk for their procedure with label-dependent stopping. Cai et al. (2023) establish a high-probability uniform estimation error bound of $\mathcal{O}(\sqrt{K\log n/n})$ under their assumptions when $K\log n/n \le 1$. Lemma 3.5 instead controls expected integrated squared estimation error directly, with a sampling term of order $K/|I|$ for $K = 2^p$. To control approximation error, we impose smoothness on the conditional class probabilities.

**Definition 3.6.** Let $\beta \in (0,1]$, $s = \kappa + \beta$, and $L > 0$. For a multi-index $\boldsymbol{\nu} = (\nu_1,\dots,\nu_d) \in \mathbb{N}_0^d$, write $|\boldsymbol{\nu}| = \sum_{j=1}^d \nu_j$ and let $\partial^{\boldsymbol{\nu}} f$ denote the corresponding mixed derivative, with $\partial^{\mathbf{0}} f = f$. A function $f : [0,1]^d \to \mathbb{R}$ belongs to the Hölder class $\mathcal{C}^{\kappa,\beta}$ of smoothness order $s$ if it has continuous mixed derivatives through order $\kappa$ and satisfies $|\partial^{\boldsymbol{\nu}} f(\mathbf{x})| \le L$ for all $|\boldsymbol{\nu}| \le \kappa$ and $\mathbf{x} \in [0,1]^d$, and $|\partial^{\boldsymbol{\nu}} f(\mathbf{x}) - \partial^{\boldsymbol{\nu}} f(\mathbf{x}')| \le L\|\mathbf{x}-\mathbf{x}'\|_2^\beta$ for all $|\boldsymbol{\nu}| = \kappa$ and $\mathbf{x},\mathbf{x}' \in [0,1]^d$.

For conditional class probabilities in $\mathcal{C}^{\kappa,\beta}$, Taylor polynomials of degree at most $\kappa$ give uniform approximation error $\mathcal{O}(\mathrm{diam}(A)^s)$ on each leaf $A$. The population polynomial projection minimizes the integrated squared error on each leaf and therefore satisfies the corresponding integrated approximation bound. Combining this with the diameter and estimation bounds yields the following expected excess risk bound.

**Theorem 3.7.** *Assume that Assumption 3.1 holds. Suppose that $\eta_c \in \mathcal{C}^{\kappa,\beta}$ for $c = 1,\dots,\tau$. Then*

$$\mathbb{E}[R(\widehat{g}_M) - R(g^*)] \le C\left((1-\alpha)^{2s\lfloor p/q\rfloor} + \frac{2^p}{n} + \lambda^2\right)^{1/2},$$

*where $C < \infty$ depends only on $(\kappa, \beta, d, \tau, L, q, \Lambda)$ and the constants implicit in $|I| \asymp n$. Moreover, choosing $p = \left\lfloor \frac{q \log n}{q \log 2 - 2s \log(1-\alpha)} \right\rfloor$ and $\lambda \asymp \sqrt{2^p/n}$ yields*

$$\mathbb{E}[R(\widehat{g}_M) - R(g^*)] \lesssim n^{\frac{s \log(1-\alpha)}{q \log 2 - 2s \log(1-\alpha)}}.$$

The chosen depth balances approximation and estimation error. For centered splits ($\alpha = 1/2$) along a basis ($q = d$), the stated tuning choices yield an expected multiclass excess risk bound of $\mathcal{O}(n^{-s/(2s+d)})$ without an additional logarithmic factor.

*Remark* 3.8. Suppose that $P_{\mathbf{X}}$ has a density $f_{\mathbf{X}}$ satisfying $0 < \underline{f} \le f_{\mathbf{X}}(\mathbf{x}) \le \overline{f} < \infty$ almost everywhere on $[0,1]^d$, with $\underline{f}$ and $\overline{f}$ fixed. The uniform design assumption in Theorem 3.7 can then be relaxed without changing the convergence exponent, provided that each leaf $A$ has an available polynomial basis orthonormal in $L_2(P_{\mathbf{X}}(\cdot \mid \mathbf{X} \in A))$, with $\phi_{A,1} \equiv 1$. The constants may additionally depend on $\underline{f}$ and $\overline{f}$.

The following theorem provides a matching lower bound on the minimax expected excess risk over Hölder classes. Let $\widetilde{g}_n$ range over all measurable, possibly randomized classifiers based on $\mathcal{D}_n$. The minimax expected excess risk is

$$R_n^* = \inf_{\widetilde{g}_n} \sup_{\eta_c \in \mathcal{C}^{\kappa,\beta}} \mathbb{E}[R(\widetilde{g}_n) - R(g^*)],$$

where the constraints on the class probabilities hold throughout $[0,1]^d$, and the expectation is taken over the training sample and any classifier randomization.

**Theorem 3.9.** *Suppose $L > 1/\tau$. For every integer $n \ge 1$, we have $R_n^* \gtrsim n^{-s/(2s+d)}$.*

Together with Theorem 3.7, this lower bound establishes that the cyclic oblique forest with centered splits ($\alpha = 1/2$) along a basis ($q = d$), and polynomial degree and tuning parameters chosen according to $s$, attains the minimax rate $n^{-s/(2s+d)}$ for expected multiclass excess risk. To the best of our knowledge, this is the first minimax guarantee without additional logarithmic factors for the expected multiclass excess risk of oblique random forests over Hölder classes of arbitrary fixed finite smoothness $s > 0$, with risk evaluated over the entire domain $[0,1]^d$ under uniform design. The analysis also extends to regression function estimation, yielding the minimax integrated mean squared error rate; details are provided in the Appendix A.5.

### 3.3 Faster Rates under a Margin Condition

We further establish margin-dependent excess risk bounds for the cyclic oblique forest over Hölder classes of any fixed finite smoothness order $s > 0$. The margin condition controls the probability that the two largest conditional class probabilities are close.

**Assumption 3.10.** *Let $\eta_{(1)}(\mathbf{x}) \ge \eta_{(2)}(\mathbf{x})$ denote the two largest conditional class probabilities. There exist fixed constants $\gamma \ge 0$ and $C_{\mathrm{margin}} > 0$ such that*

$$P_{\mathbf{X}}\{\eta_{(1)}(\mathbf{x}) - \eta_{(2)}(\mathbf{x}) \le t\} \le C_{\mathrm{margin}} t^{\gamma}, \qquad t > 0.$$

When $\gamma = 0$, the condition holds automatically with $C_{\mathrm{margin}} = 1$.

**Theorem 3.11.** *Assume that Assumptions 3.1 and 3.10 hold. Suppose that $\eta_c \in \mathcal{C}^{\kappa,\beta}$ for $c = 1, \dots, \tau$. Then*

$$\mathbb{E}[R(\widehat{g}_M) - R(g^*)] \le C_\gamma \left( (1-\alpha)^{2s\lfloor p/q \rfloor} + \frac{2^p}{n} + \lambda^2 \right)^{\frac{1+\gamma}{2+\gamma}},$$

*where $C_\gamma < \infty$ depends only on $(\kappa, \beta, d, \tau, L, q, \Lambda, \gamma, C_{\mathrm{margin}})$ and the constants implicit in $|I| \asymp n$. Moreover, choosing $p = \left\lfloor \frac{q \log n}{q \log 2 - 2s \log(1-\alpha)} \right\rfloor$ and $\lambda \asymp \sqrt{2^p/n}$ yields*

$$\mathbb{E}[R(\widehat{g}_M) - R(g^*)] \lesssim n^{\frac{2s(1+\gamma)\log(1-\alpha)}{(2+\gamma)(q \log 2 - 2s \log(1-\alpha))}}.$$

For centered splits ($\alpha = 1/2$) along a basis of $q = d$ directions, the bound gives $\mathcal{O}(n^{-2s(1+\gamma)/((2s+d)(2+\gamma))})$ without additional logarithmic factors. This recovers the no-margin rate

when $\gamma = 0$ and gives a faster rate when $\gamma > 0$. The improvement follows from the multiclass $L_2$ plug-in comparison, which changes the exponent applied to the integrated squared score error from $1/2$ to $(1+\gamma)/(2+\gamma)$. Applying this comparison to each tree and transferring the resulting excess risk bound to the majority vote yields the forest guarantee. No additional lower bound on leaf probabilities is required.

## 4 Experiments

We examine partition geometry and polynomial leaf estimation in controlled simulations, then evaluate predictive performance on real data.

### 4.1 Controlled Simulations

All simulations use centered splits ($\alpha = 1/2$) on $[0,1]^2$. For classification, we draw $\mathbf{X} \sim \mathrm{Unif}([0,1]^2)$ with

$$\eta_1(\mathbf{x}) = 0.5 + 0.4\sin(4\pi \mathbf{u}^\top \mathbf{z})\sin(4\pi \mathbf{v}^\top \mathbf{z}), \qquad \eta_2(\mathbf{x}) = 1 - \eta_1(\mathbf{x}),$$

where $\mathbf{z} = \mathbf{x} - (1/2, 1/2)^\top$, $\mathbf{u} = (\sqrt{3}/2, 1/2)^\top$, and $\mathbf{v} = (-1/2, \sqrt{3}/2)^\top$. Known class probabilities allow numerical evaluation of approximation error and excess risk.

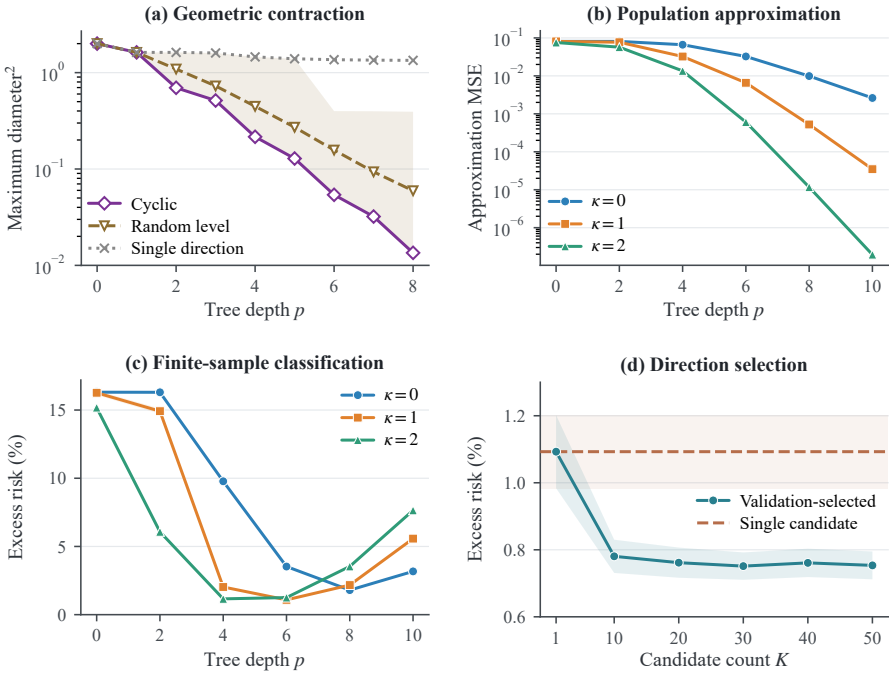


Figure 2: **Controlled simulations.** (a) Maximum squared leaf diameter. (b) Population score approximation MSE. (c) Mean excess risk of single trees. (d) Mean excess risk with validation-based direction selection. Shading shows pointwise central 95% ranges across schedules in (a) and pointwise 95% confidence intervals for the mean in (d).

Figure 2(a) compares cycling through $\mathbf{u}, \mathbf{v}$, selecting between them independently and uniformly at each depth, and repeatedly using one direction. All cells at a given depth share the selected direction. At larger depths, cyclic splitting yields smaller maximum squared leaf diameters than the average across random schedules, while the single-direction curve levels off. Figure 2(b) shows that higher polynomial degrees reduce population approximation MSE on shared cyclic partitions.

For finite-sample evaluation, we fit single trees to 100 independent samples of size $n = 10{,}000$, sharing samples and partitions across degrees. As shown in Figure 2(c), the lowest mean excess risks among the tested depths occur at $p = 8, 6, 4$ for $\kappa = 0, 1, 2$, respectively. Beyond these depths, excess

risk increases despite decreasing approximation error, illustrating the approximation–estimation trade-off.

We also evaluate validation-based direction selection using 200-tree forests with $p = 4$ and $\kappa = 2$. Each of 100 repetitions uses 7,500 training and 2,500 validation observations. Validation error selects an orthonormal basis from nested pools of $K \in \{1, 10, 20, 30, 40, 50\}$ random candidates, with subsample indices and direction-index permutations shared across candidates. The selected forest and single-basis baseline are then refitted on all 10,000 observations. Figure 2(d) shows that mean excess risk decreases from $1.09\%$ for the baseline to $0.78\%$ at $K = 10$, with limited further improvement for larger pools.

### 4.2 Predictive Performance on Real Data

We use Iranian Churn, Dry Bean, and Letter Recognition for classification, and Carbon Nanotubes for regression (UCI Machine Learning Repository, 2020b;a; Slate, 1991; UCI Machine Learning Repository, 2016). The regression task predicts $u'$ from $(n, m, u, v, w)$ and is denoted by Carbon-$u$. Baselines are SRF (Gao et al., 2022), RF (Breiman, 2001), and RRRF (Blaser & Fryzlewicz, 2016); regression comparisons use RF and RRRF.

We evaluate cyclic oblique forests with $\kappa = 0, 1, 2$ and a cyclic axis-aligned forest with $\kappa = 2$. Oblique variants select among ten fixed random orthonormal bases by validation. All forests contain 200 trees and use 100 shared holdouts with approximately $60\%$ training, $20\%$ validation, and $20\%$ test observations. Carbon-$u$ holdouts are grouped by chiral pair $(n, m)$. Model selection excludes test data. Classification depths satisfy $p < q$, so contraction over completed cycles is assessed separately in Figure 2(a).

Table 2: Test performance (mean $\pm$ sample standard deviation) over 100 shared holdouts, grouped by chiral pair for Carbon-$u$. Bold marks the lowest mean in each column; "–" denotes a method not evaluated on that task.

| | Classification error (%) | | | MSE ($\times 10^{-5}$) |
|---|---|---|---|---|
| Method | Dry Bean | Iranian Churn | Letter | Carbon-$u$ |
| Simplified RF (SRF) | $18.206 \pm 0.775$ | $9.398 \pm 1.017$ | $50.538 \pm 0.739$ | – |
| Breiman RF | $7.550 \pm 0.520$ | $4.292 \pm 0.708$ | $3.492 \pm 0.296$ | $1.333 \pm 0.554$ |
| Random rotation RF | $7.204 \pm 0.467$ | $4.384 \pm 0.751$ | $4.077 \pm 0.344$ | $14.927 \pm 6.238$ |
| Cyclic oblique ($\kappa = 0$) | $17.079 \pm 0.814$ | $8.922 \pm 0.958$ | $42.529 \pm 0.843$ | $611.031 \pm 182.992$ |
| Cyclic oblique ($\kappa = 1$) | $7.861 \pm 0.522$ | $4.510 \pm 0.790$ | $7.810 \pm 0.482$ | $1.191 \pm 0.572$ |
| Cyclic oblique ($\kappa = 2$) | $\mathbf{7.195} \pm 0.492$ | $\mathbf{3.549} \pm 0.735$ | $\mathbf{3.304} \pm 0.338$ | $\mathbf{1.145} \pm 0.546$ |
| Cyclic axis ($\kappa = 2$) | $7.465 \pm 0.492$ | $4.048 \pm 0.704$ | $4.794 \pm 0.345$ | $2.121 \pm 4.534$ |

Table 2 shows that the quadratic cyclic oblique forest achieves the lowest mean test error on all four tasks. Within the oblique family, mean error decreases with polynomial degree, and the quadratic oblique variant also achieves lower mean error than its axis-aligned counterpart across all tasks. These comparisons concern complete models, including validation-based direction selection. The shared-partition simulations provide a more direct assessment of polynomial degree.

## 5 Discussion

Future work will investigate forest constructions that attain minimax-optimal rates under joint smoothness and margin conditions. The current margin-dependent bound relies on an $L_2$ comparison and does not establish minimax optimality. Refining the splitting rules and leaf estimators may provide sharper control of class-score estimation errors and yield faster excess risk rates that match the corresponding lower bounds. In addition, learning unit normal vectors for the splitting hyperplanes from data would allow partition geometry to adapt to the classification problem. Establishing minimax guarantees for these data-dependent oblique forests, while accounting for the dependence between partition selection and statistical estimation, is worth further study.

## AI USE STATEMENT

In this work, we used generative AI tools for text drafting, grammar checking, rephrasing, and linguistic refinement of manuscript passages. We have not used generative AI tools for theorem derivation, experimental dataset generation, or the design of core algorithmic ideas, and other required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. Specifically, all AI-polished text and structural suggestions were manually inspected and revised by the authors to guarantee technical accuracy and consistency. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

## REFERENCES


Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias, 2014. URL https://arxiv.org/abs/1407.3939.

Jérémy Bettinger, François Portier, and Adrien Saumard. Pointwise convergence of purely random partition estimators: from random trees to prototype rules, 2026. URL https://arxiv.org/abs/2608.08360.

Gérard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13: 1063–1095, 2012.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(66):2015–2033, 2008. URL https://jmlr.org/papers/v9/biau08a.html.

Rico Blaser and Piotr Fryzlewicz. Random rotation ensembles. *Journal of Machine Learning Research*, 17(4):1–26, 2016. URL https://www.jmlr.org/papers/v17/blaser16a.html.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Yuchao Cai, Yuheng Ma, Yiwei Dong, and Hanfang Yang. Extrapolated random tree for regression. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3442–3468. PMLR, 2023. URL https://proceedings.mlr.press/v202/cai23d.html.

Anthony Carbery and James Wright. Distributional and $L^q$ norm inequalities for polynomials over convex bodies in $\mathbb{R}^n$. *Mathematical Research Letters*, 8(3):233–248, 2001.

Matias D. Cattaneo, Jason M. Klusowski, and William G. Underwood. Inference with Mondrian random forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(3): 1060–1085, 2026. doi: 10.1093/jrsssb/qkaf077.

Xuhui Fan, Bin Li, and Scott SIsson. Online binary space partitioning forests. In *International Conference on Artificial Intelligence and Statistics*, pp. 527–537. PMLR, 2020.

Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2):503–517, 2020.

Wei Gao and Zhi-Hua Zhou. Towards convergence rate analysis of random forests for classification. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Wei Gao, Fan Xu, and Zhi-Hua Zhou. Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103788, 2022. doi: 10.1016/j.artint.2022.103788.

Robin Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012. doi: 10.1080/10485252.2012.677843. URL https://www.tandfonline.com/doi/abs/10.1080/10485252.2012.677843.

Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R Sveinsson. Random forests for land cover classification. *Pattern recognition letters*, 27(4):294–300, 2006.

Isidoros Iakovidis and Nicola Arcozzi. Improved convergence rates for some kernel random forest algorithms, 2023. URL https://arxiv.org/abs/2310.06760.

Luckyson Khaidem, Snehanshu Saha, and Sudeepa Roy Dey. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*, 2016.

Jason Klusowski. Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pp. 757–765. PMLR, 2021.

Balaji Lakshminarayanan, Daniel M. Roy, and Yee Whye Teh. Mondrian forests: Efficient online random forests. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://arxiv.org/abs/1406.2673.

Bjoern H Menze, B Michael Kelm, Daniel N Splitthoff, Ullrich Koethe, and Fred A Hamprecht. On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 453–469. Springer, 2011.

Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Universal consistency and minimax rates for online Mondrian forests. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for Mondrian trees and forests. *The Annals of Statistics*, 48(4):2253–2276, 2020. doi: 10.1214/19-AOS1886.

Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. AMF: Aggregated Mondrian forests for online learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(3):505–533, 2021. URL https://jaouadmourtada.github.io/files/papers/mondrian_online.pdf.

Sreerama K Murthy, Simon Kasif, and Steven Salzberg. A system for induction of oblique decision trees. *Journal of artificial intelligence research*, 2:1–32, 1994.

Natalie Neumeyer, Jan Rabe, and Mathias Trabs. Asymptotic confidence bands for centered purely random forests, 2025. URL https://arxiv.org/abs/2511.13199.

Eliza O'Reilly. Statistical advantages of oblique randomized decision trees and forests, 2024. URL https://arxiv.org/abs/2407.02458. Version 3, updated 2026.

Eliza O'Reilly and Ngoc Mai Tran. Stochastic geometry to generalize the Mondrian process. *SIAM Journal on Mathematics of Data Science*, 4(2):531–552, 2022.

Eliza O'Reilly and Ngoc Mai Tran. Minimax rates for high-dimensional random tessellation forests. *Journal of Machine Learning Research*, 25(89):1–32, 2024.

Calvin Osborne and Eliza O'Reilly. The uniformly rotated Mondrian kernel. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4663–4671. PMLR, 2025. URL https://proceedings.mlr.press/v258/osborne25a.html.

Yanjun Qi. Random forest for bioinformatics. In *Ensemble machine learning*, pp. 307–323. Springer, 2012.

Tom Rainforth and Frank Wood. Canonical correlation forests. *arXiv preprint arXiv:1507.05444*, 2015.

Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016. doi: 10.1109/TIT.2016.2514489. URL https://arxiv.org/abs/1502.03836.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. 2015.

David Slate. Letter recognition dataset. UCI Machine Learning Repository, 1991. URL https://archive.ics.uci.edu/dataset/59/letter+recognition.

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

Tyler M Tomita, James Browne, Cencheng Shen, Jaewon Chung, Jesse L Patsolic, Benjamin Falk, Carey E Priebe, Jason Yim, Randal Burns, Mauro Maggioni, et al. Sparse projection oblique randomer forests. *Journal of machine learning research*, 21(104):1–39, 2020.

UCI Machine Learning Repository. Carbon nanotubes. UCI Machine Learning Repository, 2016. URL https://archive.ics.uci.edu/dataset/448/carbon+nanotubes.

UCI Machine Learning Repository. Dry bean dataset. UCI Machine Learning Repository, 2020a. URL https://archive.ics.uci.edu/dataset/602/dry+bean+dataset.

UCI Machine Learning Repository. Iranian churn dataset. UCI Machine Learning Repository, 2020b. URL https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset.

Haoran Zhan, Jingli Wang, and Yingcun Xia. Non-asymptotic properties of generalized Mondrian forests in statistical learning, 2024. URL https://arxiv.org/abs/2406.00660.

Yuqian Zhang, Weijie Ji, and Jelena Bradic. Adaptive split balancing for optimal random forest. *arXiv preprint arXiv:2402.11228*, 2024.

# A APPENDIX

## A.1 CYCLIC OBLIQUE FOREST ALGORITHM

**Algorithm 1** Cyclic Oblique Forest

**Require:** Training sample $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ and unit directions $\mathbf{b}_1, \ldots, \mathbf{b}_q$ with $M \geq 1, p \in \mathbb{N}_0, \alpha \in (0, 1/2], \kappa \in \mathbb{N}_0, \lambda > 0$ and $|I| \asymp n$.

All draws are independent of each other and of $\mathcal{D}_n$; all ties use a common deterministic rule.

**for** $m = 1, \ldots, M$ **do**
  Draw a uniform subset $I \subseteq \{1, \ldots, n\}$ of size $|I|$ and a uniform permutation $T$ of $\{1, \ldots, q\}$.
  Set $\Theta_m \leftarrow (I, T)$ and $\Pi_0 \leftarrow \{[0,1]^d\}$.
  **for** $t = 1, \ldots, p$ **do**
    Set $\Pi_t \leftarrow \varnothing$ and $j_t \leftarrow T(1 + ((t-1) \bmod q))$.
    **for** each cell $K \in \Pi_{t-1}$ **do**
      Set $m_{j_t}(K) \leftarrow \alpha \sup_{\mathbf{x} \in K} \mathbf{b}_{j_t}^\top \mathbf{x} + (1-\alpha) \inf_{\mathbf{x} \in K} \mathbf{b}_{j_t}^\top \mathbf{x}$.
      Split $K$ at $m_{j_t}(K)$ into $K^{\mathrm{L}}$ and $K^{\mathrm{R}}$ according to (2).
      Set $\Pi_t \leftarrow \Pi_t \cup \{K^{\mathrm{L}}, K^{\mathrm{R}}\}$.
    **end for**
  **end for**
  **for** each leaf $A \in \Pi_p$ **do**
    Set $I_A \leftarrow \{i \in I : \mathbf{X}_i \in A\}$.
    **if** $|I_A| > 0$ **then**
      Construct the basis vector $\boldsymbol{\phi}_A$ satisfying (3).
      Compute $\widehat{\boldsymbol{\omega}}_{A,c}$ for $c = 1, \ldots, \tau$ by (4) and $\widehat{\Omega}_A$ by (5).
      Set $\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta_m) \leftarrow \widehat{\Omega}_A^\top \boldsymbol{\phi}_A(\mathbf{x})$ for all $\mathbf{x} \in A$.
    **else**
      Set $\widehat{\eta}_c(\mathbf{x}, \Theta_m) \leftarrow |I|^{-1} \sum_{i \in I} \mathbf{1}\{Y_i = c\}$ for all $c = 1, \ldots, \tau$ and $\mathbf{x} \in A$.
    **end if**
  **end for**
  Compute $\widehat{g}(\mathbf{x}, \Theta_m) \in \arg\max_{1 \leq c \leq \tau} \widehat{\eta}_c(\mathbf{x}, \Theta_m)$.
**end for**
**return** $\widehat{g}_M$ given by (6).

## A.2 EXPERIMENTAL DETAILS

For the simulations described in Section 4.1, we set $\lambda = 0.1\sqrt{2^p/N}$, where $N$ is the sample size at the current fitting stage before tree-level subsampling.

In Figure 2(a), depths range from $p = 0$ to 8. The cyclic and single-direction curves average over the two starting directions and the two fixed directions, respectively. The random-level curve averages 200 independent direction sequences, using prefixes of the same sequences across depths. Maximum squared leaf diameters are computed from polygon vertices.

Panels (b,c) use cyclic partitions at $p \in \{0, 2, 4, 6, 8, 10\}$. The population approximation error in panel (b), $\|\boldsymbol{\eta}_{\Pi_p,\kappa} - \boldsymbol{\eta}\|_{L_2}^2$, is averaged over the two starting directions. In panel (c), each tree is fitted to the full sample, so $N = n$. Within each repetition, one uniformly sampled starting direction is shared across degrees and depths.

For panel (d), the training–validation split is generated by a label-independent random permutation. The orientations of candidate direction sets are sampled as $\theta_j \sim \mathrm{Unif}([0, \pi/2))$, independently of the data, and the first $K$ candidates form each pool. Each tree uses an $80\%$ subsample without replacement and a uniformly sampled starting direction; corresponding trees share these choices across candidates. Validation ties are resolved in favor of the smallest candidate index. Refitting uses fresh tree randomization shared across the compared forests. The fitting sample size is $N = 7{,}500$ during selection and $N = 10{,}000$ during refitting. The first candidate provides the single-candidate baseline, which coincides with the validation-selected forest at $K = 1$. Ties in tree scores or forest votes are resolved in favor of class 1. The candidate direction sets and the Sobol scrambling are regenerated independently across repetitions.

Population projections and approximation errors are computed by order-40 Gaussian quadrature on triangulated leaves, with order-20 coefficients providing a numerical check. A separate quadrature

rule checks each population Gram matrix against the identity with an operator-norm tolerance of $10^{-8}$. For a fitted classifier $\widehat{g}$, we evaluate

$$R(\widehat{g}) - R(g^*) = \int_{[0,1]^2} \big[ \max_{c \in \{1,2\}} \eta_c(\mathbf{x}) - \eta_{\widehat{g}(\mathbf{x})}(\mathbf{x}) \big] \, d\mathbf{x}$$

using $2^{16}$ scrambled Sobol points, independent of the labeled data and shared across compared models within each repetition. Neither these risk estimates nor the known probabilities enter direction selection. Excess risks are reported in percent. Confidence intervals for the mean in panel (d) are computed from the Student-$t$ distribution with 99 degrees of freedom. All summaries are computed before applying logarithmic display scales.

For the real-data experiments in Section 4.2, all cyclic oblique and axis-aligned trees start from $[0,1]^d$ and use midpoint cuts with $\alpha = 1/2$. For oblique trees, linear programming computes the projected minimum and maximum over each cell's accumulated half-space constraints. Axis-aligned trees obtain these extrema from the coordinate bounds. Each tree draws one permutation of the directions and reuses it cyclically throughout the complete depth-$p$ partition, with all nodes at a given depth sharing the scheduled direction.

All methods use the same dataset-specific preprocessing. Dry Bean and Iranian Churn use featurewise min–max transforms, with transformed values clipped to $[0,1]$ and features constant in the fitting data mapped to zero. Neither dataset requires imputation. Letter's integer-valued features are divided by their documented upper bound of 15. For Carbon-$u$, all regression methods use the same feature scaling and response standardization; test MSE is computed on the original response scale.

Classification splits are generated by label-independent random permutations. For Carbon-$u$, all original rows are retained, and the 42 chiral-pair groups are assigned intact to the three blocks to approximate the target split proportions by row count.

Each cyclic oblique or axis-aligned tree samples $80\%$ of the fitting observations without replacement. SRF, RF, and RRRF follow their native sampling rules. Candidate configurations for the cyclic variants, SRF, and RRRF are fitted on the training block and selected by validation error. RF candidates are fitted on the combined training and validation blocks and selected by out-of-bag error. Each selected configuration is then refitted on the combined blocks with a fresh seed and evaluated once on the test block. At each fitting stage, including the final refit, preprocessing parameters are estimated only from the fitting observations and held fixed for validation or test evaluation.

Table 3: Hyperparameter grids used in the real-data experiments. *Oblique* denotes the cyclic oblique variants. *Axis* denotes the axis-aligned $\kappa = 2$ control.

| Dataset | Method | Candidate values |
|---|---|---|
| Dry Bean | Oblique $\kappa = 0$ | $p \in \{6, \dots, 11\}$; $c_\lambda = 10^{-5}$ |
| | Oblique $\kappa = 1, 2$; axis | $p \in \{6, \dots, 11\}$; $c_\lambda \in \{10^{-7}, 10^{-6.5}, \dots, 10^{-5}\}$ |
| | RF | $m_{\text{try}} \in \{1, 2, 4, 8, 16\}$; $n_{\min} \in \{1, 2, 5, 10, 20\}$ |
| | RRRF | $m_{\text{try}} \in \{1, 2, 4, 10, 16\}$; $n_{\min} \in \{1, 5, 10, 15, 20\}$ |
| | SRF | $k \in \{100, 200, \dots, 2000\}$ |
| Iranian Churn | Oblique $\kappa = 0$ | $p \in \{5, 6, 7, 8\}$; $c_\lambda = 10^{-1}$ |
| | Oblique $\kappa = 1, 2$; axis | $p \in \{3, \dots, 10\}$; $c_\lambda \in \{10^{-7}, 10^{-6}, \dots, 10^{-1}\}$ |
| | RF | $m_{\text{try}} \in \{1, 2, 4, 8, 13\}$; $n_{\min} \in \{1, 2, 5, 10, 20\}$ |
| | RRRF | $m_{\text{try}} \in \{1, 2, 4, 8, 13\}$; $n_{\min} \in \{1, 2, 5, 10, 20\}$ |
| | SRF | $k \in \{50, 100, \dots, 350\}$ |
| Letter | Oblique $\kappa = 0$ | $p \in \{5, \dots, 10\}$; $c_\lambda = 10^{-1}$ |
| | Oblique $\kappa = 1, 2$; axis | $p \in \{5, \dots, 10\}$; $c_\lambda \in \{10^{-4}, 10^{-3.5}, \dots, 10^{-1}\}$ |
| | RF | $m_{\text{try}} \in \{1, 2, 4, 8, 16\}$; $n_{\min} \in \{1, 2, 5, 10, 20\}$ |
| | RRRF | $m_{\text{try}} \in \{1, 2, 4, 8, 16\}$; $n_{\min} \in \{1, 2, 5, 10, 20\}$ |
| | SRF | $k \in \{100, 200, \dots, 2000\}$ |
| Carbon-$u$ | Oblique $\kappa = 0$ | $p \in \{8, 9, 10, 11, 12\}$; $c_\lambda = 10^{-1}$ |
| | Oblique $\kappa = 1, 2$; axis | $p \in \{1, 2, 3, 4, 5\}$; $c_\lambda \in \{10^{-3}, 10^{-2}, \dots, 10^{-1}\}$ |
| | RF | $m_{\text{try}} \in \{1, 2, 4\}$; $n_{\min} \in \{1, 2, 5, 10, 20\}$ |
| | RRRF | $m_{\text{try}} \in \{1, 2, 4\}$; $n_{\min} \in \{1, 2, 5, 10, 20\}$ |

For each dataset and oblique variant, candidate sets of orthonormal directions are generated by sign-normalized QR decompositions of Gaussian matrices. The candidate library is generated before the holdouts, reused across repeats, and excludes the set of coordinate directions. The axis-aligned control shares the remaining fitting settings and depth and ridge grids of the quadratic oblique variant.

For the cyclic variants, we set $\lambda = c_\lambda \sqrt{2^p/N}$, where $c_\lambda$ is the ridge multiplier. Table 3 lists the search grids. Here $m_{\mathrm{try}}$ is the number of candidate features at an RF or RRRF split, $n_{\min}$ is the minimum leaf size, and $k$ is the maximum number of terminal leaves in SRF.

Table 4 reports paired comparisons between the quadratic oblique and axis-aligned variants within each holdout.

Table 4: Paired direction comparisons for $\kappa = 2$ over 100 shared holdouts. Reductions are axis error minus oblique error (mean $\pm$ sample standard deviation), so positive values favor the oblique variant. Classification differences are in percentage points; Carbon-$u$ differences are in units of $10^{-5}$ MSE. The final column counts wins for the oblique variant, wins for the axis-aligned variant, and ties.

| Task | Paired reduction | Oblique / axis / ties |
|---|---|---|
| Dry Bean | $0.270 \pm 0.229$ | 86/12/2 |
| Iranian Churn | $0.498 \pm 0.565$ | 78/16/6 |
| Letter | $1.490 \pm 0.366$ | 100/0/0 |
| Carbon-$u$ | $0.975 \pm 4.381$ | 64/36/0 |

### A.3 Auxiliary Lemmas

Write $p_A = P_{\mathbf{X}}(A) = \int_A d\mathbf{x}$. Conditional on any realization of $\Theta = (I, T)$, $|I_A| \sim \mathrm{Bin}(|I|, p_A)$, since the tree randomization is independent of the training sample. Throughout this subsection, expectations are conditional on $\Theta$, with this conditioning suppressed in the notation.

**Lemma A.1.** *For every leaf $A \in \Pi_p(\Theta)$,*

$$\mathbb{E}\left[\frac{\mathbf{1}\{|I_A| > 0\}}{|I_A|}\right] \le \frac{2}{(|I|+1)p_A}, \tag{7}$$

*where the ratio is defined as zero when $|I_A| = 0$. Moreover, for every $a > 0$,*

$$p_A \mathbb{E}\left[e^{-a|I_A|}\right] \le \frac{1}{e|I|(1-e^{-a})}, \tag{8}$$

**Lemma A.2.** *There exists a finite constant $L_{\mathrm{lev}}$, depending only on $(\kappa, d)$, such that*

$$\sup_{\mathbf{x}\in A} \|\boldsymbol{\phi}_A(\mathbf{x})\|_2^2 = \sup_{\mathbf{x}\in A} \sum_{j=1}^{D} \phi_{A,j}(\mathbf{x})^2 \le L_{\mathrm{lev}} \tag{9}$$

*for every leaf $A \in \Pi_p(\Theta)$ and every basis of $\mathcal{P}_\kappa(A)$ satisfying the normalized orthonormality condition above.*

**Lemma A.3.** *Let $A \in \Pi_p(\Theta)$ satisfy $|I_A| > 0$, and define $\widehat{G}_A = |I_A|^{-1} \sum_{i\in I_A} \boldsymbol{\phi}_A(\mathbf{X}_i)\boldsymbol{\phi}_A(\mathbf{X}_i)^\top$. For every $\lambda > 0$, the matrix $\widehat{G}_A + \lambda J$ is positive definite, and the unique minimizer in* (4) *is*

$$\widehat{\boldsymbol{\omega}}_{A,c} = \left(\widehat{G}_A + \lambda J\right)^{-1} \frac{1}{|I_A|} \sum_{i\in I_A} \mathbf{1}\{Y_i = c\}\boldsymbol{\phi}_A(\mathbf{X}_i), \qquad c = 1, \dots, \tau. \tag{10}$$

**Lemma A.4.** *Let $\Pi_p = \Pi_p(\Theta)$. For every $A \in \Pi_p$, $\|\Omega_A\|_F \le 1$. If $|I_A| > 0$, write $\widetilde{\Omega}_A = [\widehat{\boldsymbol{\omega}}_{A,1}, \dots, \widehat{\boldsymbol{\omega}}_{A,\tau}]$. Then*

$$\begin{aligned} \frac{1}{p_A}\int_A \left\|\widehat{\boldsymbol{\eta}}(\mathbf{x},\Theta) - \boldsymbol{\eta}_{\Pi_p}(\mathbf{x},\Theta)\right\|_2^2 d\mathbf{x} &= \|\widehat{\Omega}_A - \Omega_A\|_F^2 \\ &\le \|\widetilde{\Omega}_A - \Omega_A\|_F^2. \end{aligned} \tag{11}$$

*For every realization of the training sample and every $A \in \Pi_p$, including empty leaves,*

$$\frac{1}{p_A}\int_A \left\|\widehat{\boldsymbol{\eta}}(\mathbf{x},\Theta) - \boldsymbol{\eta}_{\Pi_p}(\mathbf{x},\Theta)\right\|_2^2 d\mathbf{x} \le 4. \tag{12}$$

**Lemma A.5.** *Let $A \in \Pi_p(\Theta)$ and $\zeta \in (0,1)$. Define $\mathcal{E}_A(\zeta) = \{|I_A| > 0, \lambda_{\min}(\widehat{G}_A) \geq \zeta\}$ and $a_\zeta = (1 - \zeta + \zeta \log \zeta)/L_{\text{lev}} > 0$, with empty leaves in $\mathcal{E}_A(\zeta)^c$. Here $\widehat{G}_A$ is defined in Lemma A.3, and $\lambda_{\min}$ denotes the smallest eigenvalue of a symmetric matrix. For every $k \in \{1, \ldots, |I|\}$ with $\mathbb{P}(|I_A| = k \mid \Theta) > 0$,*

$$\mathbb{P}(\mathcal{E}_A(\zeta)^c \,|\, \Theta, |I_A| = k) \leq D e^{-a_\zeta k}. \tag{13}$$

**Lemma A.6.** *Let $A \in \Pi_p(\Theta)$ and $\zeta \in (0,1)$. With $\mathcal{E}_A(\zeta)$ and $a_\zeta$ defined in Lemma A.5,*

$$p_A \mathbb{P}(\mathcal{E}_A(\zeta)^c \mid \Theta) \leq \frac{D}{e|I|(1 - e^{-a_\zeta})}. \tag{14}$$

**Lemma A.7.** *Suppose that $\eta_c \in \mathcal{C}^{\kappa,\beta}$ for $c = 1, \ldots, \tau$. Let $\Pi$ be a deterministic finite measurable partition of $[0,1]^d$, with $p_A = P_{\mathbf{X}}(A) > 0$ for every $A \in \Pi$. Define $\boldsymbol{\eta}_\Pi(\mathbf{x}) = \sum_{A \in \Pi} \Omega_A^\top \boldsymbol{\phi}_A(\mathbf{x}) \mathbf{1}\{\mathbf{x} \in A\}$, where the bases and coefficients are defined as in Section 2. Then*

$$\|\boldsymbol{\eta} - \boldsymbol{\eta}_\Pi\|_{L_2}^2 \leq K_{\text{app}} L^2 \sup_{A \in \Pi} \operatorname{diam}(A)^{2s}, \tag{15}$$

*where $K_{\text{app}} < \infty$ depends only on $(\kappa, \beta, d, \tau)$.*

**Lemma A.8.** *Let $\widetilde{\boldsymbol{\eta}} : [0,1]^d \to \mathbb{R}^\tau$ be measurable, with components $\widetilde{\eta}_1, \ldots, \widetilde{\eta}_\tau$, and define $\widetilde{g}(\mathbf{x}) \in \arg\max_{1 \leq c \leq \tau} \widetilde{\eta}_c(\mathbf{x})$ using the fixed deterministic tie-breaking rule. Then*

$$\begin{aligned} R(\widetilde{g}) - R(g^*) &\leq 2 \int_{[0,1]^d} \max_{1 \leq c \leq \tau} |\widetilde{\eta}_c(\mathbf{x}) - \eta_c(\mathbf{x})| \, dP_{\mathbf{X}}(\mathbf{x}) \\ &\leq 2 \int_{[0,1]^d} \|\widetilde{\boldsymbol{\eta}}(\mathbf{x}) - \boldsymbol{\eta}(\mathbf{x})\|_2 \, dP_{\mathbf{X}}(\mathbf{x}). \end{aligned} \tag{16}$$

*Under the uniform design,*

$$R(\widetilde{g}) - R(g^*) \leq 2 \|\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_{L_2}. \tag{17}$$

**Lemma A.9.** *For every nonempty bounded set $K \subset \mathbb{R}^d$, define $w_j(K) = \sup_{\mathbf{x} \in K} \mathbf{b}_j^\top \mathbf{x} - \inf_{\mathbf{x} \in K} \mathbf{b}_j^\top \mathbf{x}$ for $j = 1, \ldots, q$. Suppose that $w_j(K) > 0$ for some $j \in \{1, \ldots, q\}$. For $\alpha \in (0, 1/2]$, let $K^{\mathrm{L}}$ and $K^{\mathrm{R}}$ be the children obtained by splitting $K$ orthogonally to $\mathbf{b}_j$ at $m_j(K)$, as in (2). Both children are nonempty, and*

$$w_j(K^{\mathrm{L}}) \leq \alpha w_j(K), \qquad w_j(K^{\mathrm{R}}) \leq (1 - \alpha) w_j(K).$$

*Moreover, $w_k(K') \leq w_k(K)$ for every nonempty $K' \subseteq K$ and every $k = 1, \ldots, q$.*

### A.4 Proofs of the Main Results

*Proof of Lemma 3.4.* For every nonempty bounded set $K \subset \mathbb{R}^d$ and $\mathbf{z}, \mathbf{z}' \in K$, Assumption 3.1 gives

$$\|\mathbf{z} - \mathbf{z}'\|_2^2 \leq \Lambda \sum_{j=1}^q \big(\mathbf{b}_j^\top (\mathbf{z} - \mathbf{z}')\big)^2 \leq \Lambda \sum_{j=1}^q w_j(K)^2.$$

Taking the supremum over $\mathbf{z}, \mathbf{z}' \in K$ yields

$$\operatorname{diam}(K)^2 \leq \Lambda \sum_{j=1}^q w_j(K)^2. \tag{18}$$

Fix $\Theta = (I, T)$. Every cell is convex, since each child is the intersection of its parent with a half-space. The root has nonempty interior. For a convex cell with nonempty interior, the splitting threshold lies strictly between the infimum and supremum of its projection. The splitting hyperplane therefore intersects its interior, and both children have nonempty interior. By induction, every cell has nonempty interior, so Lemma A.9 applies at every split.

Fix $\mathbf{x} \in [0,1]^d$. Along the path to $A_p(\mathbf{x})$, each direction is selected at least $\lfloor p/q \rfloor$ times. Since $\alpha \leq 1 - \alpha$, each split in direction $\mathbf{b}_j$ contracts $w_j$ by a factor at most $1 - \alpha$, while all other splits cannot increase it. Consequently,

$$w_j(A_p(\mathbf{x})) \leq (1 - \alpha)^{\lfloor p/q \rfloor} w_j([0,1]^d), \qquad j = 1, \ldots, q.$$

By Cauchy–Schwarz and $\|\mathbf{b}_j\|_2 = 1$, $w_j([0,1]^d) \le \operatorname{diam}([0,1]^d) = \sqrt{d}$. Applying (18) therefore gives

$$\begin{aligned}\operatorname{diam}(A_p(\mathbf{x}))^2 &\le \Lambda(1-\alpha)^{2\lfloor p/q \rfloor} \sum_{j=1}^{q} w_j([0,1]^d)^2 \\ &\le \Lambda q d (1-\alpha)^{2\lfloor p/q \rfloor}.\end{aligned}$$

The bound holds for every $\Theta$ and $\mathbf{x}$. □

*Proof of Lemma 3.5.* Consider a realization of $\Theta = (I, T)$ and write $\Pi_p = \Pi_p(\Theta)$, so that $|\Pi_p| = 2^p$. The observations indexed by $I$ remain i.i.d. with law $P$, since $\Theta$ is independent of the training sample. Unless stated otherwise, all expectations and probabilities below are conditional on $\Theta$.

For $A \in \Pi_p$ and $k \in \{1, \dots, |I|\}$ with $\mathbb{P}(|I_A| = k) > 0$, list the observations in $A$ as $(\mathbf{X}_{A,1}, Y_{A,1}), \dots, (\mathbf{X}_{A,k}, Y_{A,k})$ in increasing order of their original sample indices. Conditional on $|I_A| = k$, they are i.i.d. from $P(\cdot \mid \mathbf{X} \in A)$. Let $\boldsymbol{\omega}_{A,c}$ denote column $c$ of $\Omega_A$, and define

$$\boldsymbol{\Delta}_{A,c} = \frac{1}{k} \sum_{r=1}^{k} \left( \mathbf{1}\{Y_{A,r} = c\} - \boldsymbol{\phi}_A(\mathbf{X}_{A,r})^\top \boldsymbol{\omega}_{A,c} \right) \boldsymbol{\phi}_A(\mathbf{X}_{A,r}), \qquad c = 1, \dots, \tau.$$

The summands are centered by the definition of $\Omega_A$ and orthonormality. Also, $\|\boldsymbol{\omega}_{A,c}\|_2^2 \le 1$ by Lemma A.4. Conditional independence, orthonormality, and Lemma A.2 give

$$\begin{aligned}\mathbb{E}\left[ \|\boldsymbol{\Delta}_{A,c}\|_2^2 \,\middle|\, |I_A| = k \right] &= \frac{1}{k} \mathbb{E}\left[ \left( \mathbf{1}\{Y = c\} - \boldsymbol{\phi}_A(\mathbf{X})^\top \boldsymbol{\omega}_{A,c} \right)^2 \|\boldsymbol{\phi}_A(\mathbf{X})\|_2^2 \,\middle|\, \mathbf{X} \in A \right] \\ &\le \frac{2L_{\mathrm{lev}}}{k} \left( 1 + \|\boldsymbol{\omega}_{A,c}\|_2^2 \right) \le \frac{4L_{\mathrm{lev}}}{k}.\end{aligned}$$

Thus, for $\Delta_A = [\boldsymbol{\Delta}_{A,1}, \dots, \boldsymbol{\Delta}_{A,\tau}]$,

$$\mathbb{E}\left[ \|\Delta_A\|_F^2 \,\middle|\, |I_A| = k \right] \le \frac{4\tau L_{\mathrm{lev}}}{k}. \tag{19}$$

Use the event $\mathcal{E}_A(1/2)$ defined in Lemma A.6. On this event, $\widehat{G}_A$ has smallest eigenvalue at least $1/2$. Since $J$ is positive semidefinite, $\|(\widehat{G}_A + \lambda J)^{-1}\|_{\mathrm{op}} \le 2$. With $\widetilde{\Omega}_A$ as in Lemma A.4, the ridge normal equations give

$$\widetilde{\Omega}_A - \Omega_A = (\widehat{G}_A + \lambda J)^{-1} (\Delta_A - \lambda J \Omega_A).$$

Since $\|J\Omega_A\|_F \le \|\Omega_A\|_F \le 1$, it follows that $\|\widetilde{\Omega}_A - \Omega_A\|_F^2 \le 8\|\Delta_A\|_F^2 + 8\lambda^2$ on $\mathcal{E}_A(1/2)$. Combining (19) with Lemma A.4 yields

$$\begin{aligned}&\mathbb{E}\left[ \frac{1}{p_A} \int_A \left\| \widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta) - \boldsymbol{\eta}_{\Pi_p}(\mathbf{x}, \Theta) \right\|_2^2 d\mathbf{x}\, \mathbf{1}\{\mathcal{E}_A(1/2)\} \,\middle|\, |I_A| = k \right] \\ &\qquad \le \frac{32\tau L_{\mathrm{lev}}}{k} + 8\lambda^2.\end{aligned} \tag{20}$$

On $\mathcal{E}_A(1/2)^c$, including empty leaves, the normalized squared score error is at most 4 by (12). Averaging (20) over the leaf count therefore gives

$$\begin{aligned}&\mathbb{E}\left[ \frac{1}{p_A} \int_A \left\| \widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta) - \boldsymbol{\eta}_{\Pi_p}(\mathbf{x}, \Theta) \right\|_2^2 d\mathbf{x} \right] \\ &\qquad \le 32\tau L_{\mathrm{lev}} \mathbb{E}\left[ \frac{\mathbf{1}\{|I_A| > 0\}}{|I_A|} \right] + 8\lambda^2 + 4\mathbb{P}(\mathcal{E}_A(1/2)^c).\end{aligned}$$

The inverse-count ratio is zero on empty leaves, as in Lemma A.1. Multiplying by $p_A$ and applying Lemmas A.1 and A.6 with $\zeta = 1/2$ gives

$$\mathbb{E}\left[ \int_A \left\| \widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta) - \boldsymbol{\eta}_{\Pi_p}(\mathbf{x}, \Theta) \right\|_2^2 d\mathbf{x} \right] \le C_{\mathrm{est}} \left( \frac{1}{|I|} + \lambda^2 p_A \right),$$

where $C_{\mathrm{est}} < \infty$ is chosen large enough to absorb the numerical coefficients and depends only on $(\kappa, d, \tau)$ and the constants implicit in $|I| \asymp n$. Summing over $A \in \Pi_p$, using $\sum_{A\in\Pi_p} p_A = 1$ and $|\Pi_p| = 2^p$, and restoring the conditioning on $\Theta$ yields

$$\mathbb{E}\left[\left\|\widehat{\boldsymbol{\eta}}(\cdot,\Theta) - \boldsymbol{\eta}_{\Pi_p}(\cdot,\Theta)\right\|_{L_2}^2 \,\middle|\, \Theta\right] \le C_{\mathrm{est}}\left(\frac{2^p}{|I|} + \lambda^2\right). \tag{21}$$

Since $|I| \asymp n$, choosing $\lambda \asymp \sqrt{2^p/n}$ in (21) gives the stated $\mathcal{O}(2^p/n)$ bound. □

*Proof of Theorem 3.7.* Condition on $\Theta = (I, T)$ and write $\Pi_p = \Pi_p(\Theta)$. The partition and subsample index set are then deterministic, and conditional expectations are over the training sample. By Lemmas A.7 and 3.4,

$$\begin{aligned}\|\boldsymbol{\eta} - \boldsymbol{\eta}_{\Pi_p}(\cdot,\Theta)\|_{L_2}^2 &\le K_{\mathrm{app}} L^2 \sup_{A\in\Pi_p} \operatorname{diam}(A)^{2s} \\ &\le C_1(1-\alpha)^{2s\lfloor p/q\rfloor},\end{aligned}$$

where $C_1 = K_{\mathrm{app}} L^2 (\Lambda q d)^s$.

On each leaf, every component of $\widehat{\boldsymbol{\eta}}(\cdot,\Theta)$ belongs to $\mathcal{P}_\kappa(A)$, including the constant prediction on an empty leaf. Since $\boldsymbol{\eta}_{\Pi_p}(\cdot,\Theta)$ is the componentwise orthogonal projection of $\boldsymbol{\eta}$, the Pythagorean identity, Lemma 3.5, and $|I| \asymp n$ give

$$\begin{aligned}&\mathbb{E}\big[\|\widehat{\boldsymbol{\eta}}(\cdot,\Theta) - \boldsymbol{\eta}\|_{L_2}^2 \,\big|\, \Theta\big] \\ &\quad= \mathbb{E}\Big[\|\widehat{\boldsymbol{\eta}}(\cdot,\Theta) - \boldsymbol{\eta}_{\Pi_p}(\cdot,\Theta)\|_{L_2}^2 \,\Big|\, \Theta\Big] \\ &\qquad + \|\boldsymbol{\eta}_{\Pi_p}(\cdot,\Theta) - \boldsymbol{\eta}\|_{L_2}^2 \\ &\quad\le C_1(1-\alpha)^{2s\lfloor p/q\rfloor} + C_2\left(\frac{2^p}{n} + \lambda^2\right),\end{aligned} \tag{22}$$

where $C_2 < \infty$ depends only on $(\kappa, d, \tau)$ and the constants implicit in $|I| \asymp n$. The same bound holds after averaging over $\Theta$.

For each $\mathbf{x}$, let $h = \widehat{g}_M(\mathbf{x})$. By (6), $\sum_{m=1}^M \mathbf{1}\{\widehat{g}(\mathbf{x},\Theta_m) = h\} \ge M/\tau$. Since $\eta_{g^*(\mathbf{x})}(\mathbf{x}) - \eta_c(\mathbf{x}) \ge 0$ for every class $c$,

$$\begin{aligned}&\frac{1}{M}\sum_{m=1}^M \left[\eta_{g^*(\mathbf{x})}(\mathbf{x}) - \eta_{\widehat{g}(\mathbf{x},\Theta_m)}(\mathbf{x})\right] \\ &\quad\ge \frac{1}{M}\sum_{m=1}^M \mathbf{1}\{\widehat{g}(\mathbf{x},\Theta_m) = h\}\left[\eta_{g^*(\mathbf{x})}(\mathbf{x}) - \eta_h(\mathbf{x})\right] \\ &\quad\ge \frac{1}{\tau}\left[\eta_{g^*(\mathbf{x})}(\mathbf{x}) - \eta_h(\mathbf{x})\right].\end{aligned}$$

Integrating with respect to $P_{\mathbf{X}}$ gives

$$R(\widehat{g}_M) - R(g^*) \le \frac{\tau}{M}\sum_{m=1}^M \left[R(\widehat{g}(\cdot,\Theta_m)) - R(g^*)\right]. \tag{23}$$

Taking expectations in (23) and applying Lemma A.8, Cauchy–Schwarz, and (22), we obtain

$$\begin{aligned}\mathbb{E}[R(\widehat{g}_M) - R(g^*)] &\le \frac{2\tau}{M}\sum_{m=1}^M \mathbb{E}[\|\widehat{\boldsymbol{\eta}}(\cdot,\Theta_m) - \boldsymbol{\eta}\|_{L_2}] \\ &\le 2\tau\left(C_1(1-\alpha)^{2s\lfloor p/q\rfloor} + C_2\left(\frac{2^p}{n} + \lambda^2\right)\right)^{1/2} \\ &\le C\left((1-\alpha)^{2s\lfloor p/q\rfloor} + \frac{2^p}{n} + \lambda^2\right)^{1/2}.\end{aligned}$$

Here $C = 2\tau\sqrt{\max\{C_1, C_2\}} < \infty$ depends only on $(\kappa, \beta, d, \tau, L, q, \Lambda)$ and the constants implicit in $|I| \asymp n$. This proves the excess risk bound in Theorem 3.7.

Choose $p = \lfloor q \log n / [q \log 2 - 2s \log(1-\alpha)] \rfloor$ and $\lambda \asymp \sqrt{2^p/n}$. Since $q$ is an integer, $\lfloor p/q \rfloor \geq \log n / [q \log 2 - 2s \log(1-\alpha)] - 1$. Therefore,

$$\frac{2^p}{n} \leq n^{\frac{2s\log(1-\alpha)}{q\log 2 - 2s\log(1-\alpha)}},$$
$$(1-\alpha)^{2s\lfloor p/q \rfloor} \leq (1-\alpha)^{-2s} n^{\frac{2s\log(1-\alpha)}{q\log 2 - 2s\log(1-\alpha)}}.$$

Substituting these bounds and $\lambda^2 \asymp 2^p/n$ into the excess risk bound just established yields

$$\mathbb{E}[R(\widehat{g}_M) - R(g^*)] \lesssim n^{\frac{s\log(1-\alpha)}{q\log 2 - 2s\log(1-\alpha)}}.$$

The implicit constant depends only on the fixed model and sampling parameters. For $\alpha = 1/2$, the exponent is $-s/(2s+q)$; when $q = d$, it is $-s/(2s+d)$. □

*Proof of Remark 3.8.* Suppose that $0 < \underline{f} \leq f_{\mathbf{X}}(\mathbf{x}) \leq \overline{f} < \infty$ almost everywhere on $[0,1]^d$. Consider a realization of $\Theta = (I, T)$ and write $\Pi_p = \Pi_p(\Theta)$. For each $A \in \Pi_p$, write $p_A = P_{\mathbf{X}}(A) = \int_A f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x}$. Use the assumed polynomial basis, orthonormal under $P_{\mathbf{X}}(\cdot \mid \mathbf{X} \in A)$, with $\phi_{A,1} \equiv 1$. The population coefficients $\Omega_A$ and projection $\boldsymbol{\eta}_{\Pi_p}(\cdot, \Theta)$ are defined using this basis and conditional distribution.

Applying (38) to the uniform measure on $A$ and using the density bounds gives, for every $r \in \mathcal{P}_\kappa(A)$,

$$\begin{aligned}\sup_{\mathbf{x}\in A} |r(\mathbf{x})|^2 &\leq \frac{L_{\mathrm{lev}}}{\int_A d\mathbf{x}} \int_A |r(\mathbf{x})|^2\, d\mathbf{x} \\ &\leq L_{\mathrm{lev}} \frac{\overline{f}}{\underline{f}} \mathbb{E}\big[|r(\mathbf{X})|^2 \mid \mathbf{X} \in A\big],\end{aligned}$$

where $L_{\mathrm{lev}}$ retains its value from Lemma A.2. The second inequality uses $\int_A |r(\mathbf{x})|^2\, d\mathbf{x} \leq (p_A/\underline{f})\mathbb{E}[|r(\mathbf{X})|^2 \mid \mathbf{X} \in A]$ and $p_A \leq \overline{f} \int_A d\mathbf{x}$. For every $\mathbf{u} \in \mathbb{R}^D$ with $\|\mathbf{u}\|_2 = 1$, orthonormality gives $\mathbb{E}[|\mathbf{u}^\top \boldsymbol{\phi}_A(\mathbf{X})|^2 \mid \mathbf{X} \in A] = 1$. Applying the preceding bound and taking the supremum over $\mathbf{u}$ yields

$$\sup_{\mathbf{x}\in A} \|\boldsymbol{\phi}_A(\mathbf{x})\|_2^2 \leq L_{\mathrm{lev}} \frac{\overline{f}}{\underline{f}}.$$

This bound is uniform over all leaves.

Given $\Theta$ and a nonzero leaf count, the leaf observations remain i.i.d. from $P(\cdot \mid \mathbf{X} \in A)$. Orthonormality keeps the population Gram matrix equal to the identity and the residuals in the preceding estimation proof centered. The Gram and residual bounds therefore apply with leverage bound $L_{\mathrm{lev}}\overline{f}/\underline{f}$. The corresponding Chernoff exponent is

$$a_\zeta^{(f)} = \frac{1 - \zeta + \zeta\log\zeta}{L_{\mathrm{lev}}\overline{f}/\underline{f}}, \qquad \zeta \in (0,1).$$

The Gram lower-tail and weighted bad-Gram probability bounds therefore hold with $a_\zeta$ replaced by $a_\zeta^{(f)}$.

Conditional orthonormality and the fact that $\boldsymbol{\eta}$ is a probability vector give $\|\Omega_A\|_F \leq 1$. Thus Lemma A.4, including its empty-leaf bound, holds with normalized Lebesgue integrals replaced by conditional expectations given $\mathbf{X} \in A$.

Since $|I_A| \mid \Theta \sim \mathrm{Bin}(|I|, p_A)$, the inverse-count and weighted bad-Gram arguments also apply. Repeating the proof of Lemma 3.5 with $\mathcal{E}_A(1/2)$ and summing with weights $p_A$ gives

$$\mathbb{E}\left[\int_{[0,1]^d} \left\|\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta) - \boldsymbol{\eta}_{\Pi_p}(\mathbf{x}, \Theta)\right\|_2^2 dP_{\mathbf{X}}(\mathbf{x}) \,\middle|\, \Theta\right] \leq C_{\mathrm{est}} \frac{2^p}{|I|} + 8\lambda^2,$$

where $C_{\mathrm{est}} < \infty$ depends only on $(\kappa, d, \tau, \underline{f}, \overline{f})$.

The Taylor remainder bound in Lemma A.7 is pointwise. Since the population polynomial projection minimizes squared error under the conditional distribution on each leaf, the same argument gives

$$\int_{[0,1]^d} \left\| \boldsymbol{\eta}(\mathbf{x}) - \boldsymbol{\eta}_{\Pi_p}(\mathbf{x}, \Theta) \right\|_2^2 \, dP_{\mathbf{X}}(\mathbf{x}) \le K_{\mathrm{app}} L^2 \sup_{A \in \Pi_p} \operatorname{diam}(A)^{2s}.$$

The deterministic diameter bound is unchanged. Moreover, (16) and Cauchy–Schwarz give

$$R(\widetilde{g}) - R(g^*) \le 2 \left( \int_{[0,1]^d} \|\widetilde{\boldsymbol{\eta}}(\mathbf{x}) - \boldsymbol{\eta}(\mathbf{x})\|_2^2 \, dP_{\mathbf{X}}(\mathbf{x}) \right)^{1/2}.$$

The majority-vote argument also applies to this feature distribution. Combining these bounds as in the proof of Theorem 3.7 gives the same risk bound and convergence exponent, with constants that may additionally depend on $\underline{f}$ and $\overline{f}$. □

*Proof of Theorem 3.9.* Choose a nonzero, nonnegative smooth function $\psi : \mathbb{R}^d \to \mathbb{R}$ with compact support in $(1/4, 3/4)^d$, depending only on $d$. For an integer $r \ge 2$, let $h = 1/r$ and $\mathcal{J}_r = \{0, \dots, r-1\}^d$. For $\varepsilon > 0$ to be chosen and $\boldsymbol{\sigma} = (\sigma_{\mathbf{j}})_{\mathbf{j} \in \mathcal{J}_r} \in \{-1, 1\}^{r^d}$, define

$$u_{\mathbf{j}}(\mathbf{x}) = \varepsilon h^s \psi\left( \frac{\mathbf{x} - h\mathbf{j}}{h} \right), \qquad F_{\boldsymbol{\sigma},h}(\mathbf{x}) = \sum_{\mathbf{j} \in \mathcal{J}_r} \sigma_{\mathbf{j}} u_{\mathbf{j}}(\mathbf{x}).$$

The supports of the bumps are disjoint and contained in $[0,1]^d$, and

$$\partial^{\boldsymbol{\nu}} u_{\mathbf{j}}(\mathbf{x}) = \varepsilon h^{s - |\boldsymbol{\nu}|} (\partial^{\boldsymbol{\nu}} \psi)\left( \frac{\mathbf{x} - h\mathbf{j}}{h} \right).$$

Since $h \le 1$ and $s = \kappa + \beta$, all derivatives through order $\kappa$ are bounded uniformly in $h$. For $|\boldsymbol{\nu}| = \kappa$, scaling preserves the global $\beta$-Hölder bound up to the factor $\varepsilon$. At any two points, at most two bumps contribute to the derivative difference. Thus there is a finite $C_\psi$, depending only on $(d, \kappa, \beta)$, such that $|\partial^{\boldsymbol{\nu}} F_{\boldsymbol{\sigma},h}(\mathbf{x})| \le \varepsilon C_\psi$ for $|\boldsymbol{\nu}| \le \kappa$, and

$$|\partial^{\boldsymbol{\nu}} F_{\boldsymbol{\sigma},h}(\mathbf{x}) - \partial^{\boldsymbol{\nu}} F_{\boldsymbol{\sigma},h}(\mathbf{y})| \le \varepsilon C_\psi \|\mathbf{x} - \mathbf{y}\|_2^{\beta}, \qquad |\boldsymbol{\nu}| = \kappa.$$

These bounds hold uniformly in $r$ and $\boldsymbol{\sigma}$.

Since $L > 1/\tau$, choose a fixed $\varepsilon > 0$ such that

$$\varepsilon C_\psi \le \min\left\{ L - \frac{1}{\tau}, \frac{1}{2\tau} \right\}, \qquad 8\tau\varepsilon^2 \int_{\mathbb{R}^d} \psi(\mathbf{z})^2 \, d\mathbf{z} \le \frac{1}{2}.$$

Define $\eta_{\boldsymbol{\sigma},1} = 1/\tau + F_{\boldsymbol{\sigma},h}$, $\eta_{\boldsymbol{\sigma},2} = 1/\tau - F_{\boldsymbol{\sigma},h}$, and $\eta_{\boldsymbol{\sigma},c} \equiv 1/\tau$ for $c = 3, \dots, \tau$. The preceding bounds imply $\eta_{\boldsymbol{\sigma},c} \in \mathcal{C}^{\kappa,\beta}$ by Definition 3.6. Moreover, the probabilities sum to one and are all at least $1/(2\tau)$. Let $P_{\boldsymbol{\sigma}}$ have feature marginal $P_{\mathbf{X}} = \mathrm{Unif}([0,1]^d)$ and these conditional class probabilities. Every $P_{\boldsymbol{\sigma}}$ is therefore admissible in the definition of $R_n^*$.

Write $R_{\boldsymbol{\sigma}}$ and $g_{\boldsymbol{\sigma}}^*$ for the risk and a Bayes classifier under $P_{\boldsymbol{\sigma}}$. Let $\widetilde{g}_n$ be any jointly measurable, possibly randomized classifier based on $\mathcal{D}_n$. Its auxiliary randomization is independent of the sample and has the same law under every model. Write $\mathbb{E}_{\boldsymbol{\sigma}}$ and $\mathbb{P}_{\boldsymbol{\sigma}}$ for expectation and probability over both the sample and this randomization. For each $\mathbf{j}$, let $c_{\boldsymbol{\sigma},\mathbf{j}} = 1$ if $\sigma_{\mathbf{j}} = 1$ and $c_{\boldsymbol{\sigma},\mathbf{j}} = 2$ otherwise. Where $u_{\mathbf{j}}(\mathbf{x}) > 0$, this is the unique Bayes class, and any other prediction incurs pointwise excess risk at least $u_{\mathbf{j}}(\mathbf{x})$. Consequently, Tonelli's theorem gives

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\sigma}}[R_{\boldsymbol{\sigma}}(\widetilde{g}_n) - R_{\boldsymbol{\sigma}}(g_{\boldsymbol{\sigma}}^*)] \\ &\quad \ge \sum_{\mathbf{j} \in \mathcal{J}_r} \int_{[0,1]^d} u_{\mathbf{j}}(\mathbf{x}) \mathbb{P}_{\boldsymbol{\sigma}}\{\widetilde{g}_n(\mathbf{x}) \ne c_{\boldsymbol{\sigma},\mathbf{j}}\} \, d\mathbf{x}. \end{aligned} \tag{24}$$

Let $\boldsymbol{\sigma}^{(\mathbf{j})}$ be obtained by flipping only $\sigma_{\mathbf{j}}$. The two models agree outside the support of $u_{\mathbf{j}}$ and interchange the probabilities $1/\tau + u_{\mathbf{j}}$ and $1/\tau - u_{\mathbf{j}}$ on that support. Write $D_{\mathrm{KL}}$ for Kullback–Leibler

divergence and $\|P-Q\|_{\mathrm{TV}} = \sup_B |P(B)-Q(B)|$, where $B$ ranges over measurable events. Since $0 \le u_{\mathbf{j}} \le 1/(2\tau)$ and $\log((1+t)/(1-t)) \le 4t$ for $0 \le t \le 1/2$,

$$\begin{aligned} D_{\mathrm{KL}}(P_{\boldsymbol{\sigma}} \| P_{\boldsymbol{\sigma}^{(\mathbf{j})}}) &\le 8\tau \int_{[0,1]^d} u_{\mathbf{j}}(\mathbf{x})^2 \, d\mathbf{x} \\ &= 8\tau\varepsilon^2 h^{2s+d} \int_{\mathbb{R}^d} \psi(\mathbf{z})^2 \, d\mathbf{z}. \end{aligned}$$

For every $n \ge 1$, choose $r = 1 + \lfloor n^{1/(2s+d)} \rfloor$. Then $nh^{2s+d} \le 1$ and $h \ge \frac{1}{2} n^{-1/(2s+d)}$. Write $P_{\boldsymbol{\sigma}}^{\otimes n}$ for the joint law of the $n$ i.i.d. training observations under $P_{\boldsymbol{\sigma}}$. By the choice of $\varepsilon$, additivity of Kullback–Leibler divergence and Pinsker's inequality yield

$$\left\| P_{\boldsymbol{\sigma}}^{\otimes n} - P_{\boldsymbol{\sigma}^{(\mathbf{j})}}^{\otimes n} \right\|_{\mathrm{TV}} \le \left[ \frac{n}{2} D_{\mathrm{KL}}(P_{\boldsymbol{\sigma}} \| P_{\boldsymbol{\sigma}^{(\mathbf{j})}}) \right]^{1/2} \le \frac{1}{2}. \tag{25}$$

The same total variation bound holds after adding the common independent classifier randomization.

For any $\mathbf{x}$ and neighboring pair with $\sigma_{\mathbf{j}} = 1$, the events $\{\widetilde{g}_n(\mathbf{x}) = 1\}$ and $\{\widetilde{g}_n(\mathbf{x}) = 2\}$ are disjoint. Thus (25) gives

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{\sigma}}\{\widetilde{g}_n(\mathbf{x}) \ne 1\} + \mathbb{P}_{\boldsymbol{\sigma}^{(\mathbf{j})}}\{\widetilde{g}_n(\mathbf{x}) \ne 2\} \\ &\quad \ge 1 - \mathbb{P}_{\boldsymbol{\sigma}}\{\widetilde{g}_n(\mathbf{x}) = 1\} + \mathbb{P}_{\boldsymbol{\sigma}^{(\mathbf{j})}}\{\widetilde{g}_n(\mathbf{x}) = 1\} \\ &\quad \ge 1 - \left\| P_{\boldsymbol{\sigma}}^{\otimes n} - P_{\boldsymbol{\sigma}^{(\mathbf{j})}}^{\otimes n} \right\|_{\mathrm{TV}} \ge \frac{1}{2}. \end{aligned}$$

Pairing the $2^{r^d}$ sign vectors by coordinate $\mathbf{j}$ therefore gives

$$2^{-r^d} \sum_{\boldsymbol{\sigma}} \mathbb{P}_{\boldsymbol{\sigma}}\{\widetilde{g}_n(\mathbf{x}) \ne c_{\boldsymbol{\sigma},\mathbf{j}}\} \ge \frac{1}{4}.$$

Averaging (24) over all sign vectors and using $|\mathcal{J}_r| = h^{-d}$, we obtain

$$\begin{aligned} &2^{-r^d} \sum_{\boldsymbol{\sigma}} \mathbb{E}_{\boldsymbol{\sigma}}[R_{\boldsymbol{\sigma}}(\widetilde{g}_n) - R_{\boldsymbol{\sigma}}(g^*_{\boldsymbol{\sigma}})] \\ &\quad \ge \frac{1}{4} \sum_{\mathbf{j} \in \mathcal{J}_r} \int_{[0,1]^d} u_{\mathbf{j}}(\mathbf{x}) \, d\mathbf{x} \\ &\quad = \frac{\varepsilon h^s}{4} \int_{\mathbb{R}^d} \psi(\mathbf{z}) \, d\mathbf{z} \gtrsim n^{-s/(2s+d)}. \end{aligned}$$

The worst-case risk over the admissible model class is at least this finite average. Taking the infimum over $\widetilde{g}_n$ proves $R_n^* \gtrsim n^{-s/(2s+d)}$ for every $n \ge 1$. The implicit constant depends only on $(d, \kappa, \beta, \tau, L)$. □

*Proof of Theorem 3.11.* Given $\mathcal{D}_n$ and $\Theta = (I, T)$, define $u(\mathbf{x}) = \max_{1 \le c \le \tau} |\widehat{\eta}_c(\mathbf{x}, \Theta) - \eta_c(\mathbf{x})|$ and $\ell(\mathbf{x}) = \eta_{g^*(\mathbf{x})}(\mathbf{x}) - \eta_{\widehat{g}(\mathbf{x},\Theta)}(\mathbf{x})$. By the argmax definition of $\widehat{g}(\mathbf{x}, \Theta)$, $0 \le \ell(\mathbf{x}) \le 2u(\mathbf{x})$. Whenever $\ell(\mathbf{x}) > 0$, the selected class is not Bayes optimal, so $\eta_{(1)}(\mathbf{x}) - \eta_{(2)}(\mathbf{x}) \le \ell(\mathbf{x}) \le 2u(\mathbf{x})$.

For every $t > 0$, splitting according to $u(\mathbf{x}) \le t$ and $u(\mathbf{x}) > t$, Assumption 3.10 and the uniform design give

$$\begin{aligned} R(\widehat{g}(\cdot, \Theta)) - R(g^*) &= \int_{[0,1]^d} \ell(\mathbf{x}) \, d\mathbf{x} \\ &\le 2t P_{\mathbf{X}}\{\eta_{(1)}(\mathbf{X}) - \eta_{(2)}(\mathbf{X}) \le 2t\} \\ &\quad + 2 \int_{[0,1]^d} u(\mathbf{x}) \mathbf{1}\{u(\mathbf{x}) > t\} \, d\mathbf{x} \\ &\le 2^{1+\gamma} C_{\mathrm{margin}} t^{1+\gamma} + \frac{2}{t} \|\widehat{\boldsymbol{\eta}}(\cdot, \Theta) - \boldsymbol{\eta}\|_{L_2}^2. \end{aligned}$$

The last inequality uses $u(\mathbf{x}) \mathbf{1}\{u(\mathbf{x}) > t\} \le u(\mathbf{x})^2/t$ and $u(\mathbf{x})^2 \le \|\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta) - \boldsymbol{\eta}(\mathbf{x})\|_2^2$. If the squared $L_2$ error is zero, the excess risk is zero. Otherwise, taking $t = \|\widehat{\boldsymbol{\eta}}(\cdot, \Theta) - \boldsymbol{\eta}\|_{L_2}^{2/(2+\gamma)}$ yields

$$R(\widehat{g}(\cdot, \Theta)) - R(g^*) \le (2^{1+\gamma} C_{\mathrm{margin}} + 2) \|\widehat{\boldsymbol{\eta}}(\cdot, \Theta) - \boldsymbol{\eta}\|_{L_2}^{\frac{2(1+\gamma)}{2+\gamma}}. \tag{26}$$

This inequality holds for every realization of the sample and tree randomization.

Since $(1+\gamma)/(2+\gamma) \in [1/2, 1)$, conditional Jensen's inequality, (26), and (22) imply

$$\begin{aligned}
&\mathbb{E}[R(\widehat{g}(\cdot,\Theta)) - R(g^*) \,|\, \Theta] \\
&\lesssim \left(\mathbb{E}\left[\|\widehat{\boldsymbol{\eta}}(\cdot,\Theta) - \boldsymbol{\eta}\|_{L_2}^2 \,\middle|\, \Theta\right]\right)^{\frac{1+\gamma}{2+\gamma}} \\
&\lesssim \left((1-\alpha)^{2s\lfloor p/q \rfloor} + \frac{2^p}{n} + \lambda^2\right)^{\frac{1+\gamma}{2+\gamma}}.
\end{aligned}$$

The bound is uniform over $\Theta$ and therefore also holds after averaging over the tree randomization.

By the voting comparison (23),

$$\begin{aligned}
\mathbb{E}[R(\widehat{g}_M) - R(g^*)] &\le \frac{\tau}{M} \sum_{m=1}^{M} \mathbb{E}[R(\widehat{g}(\cdot,\Theta_m)) - R(g^*)] \\
&\le C_\gamma \left((1-\alpha)^{2s\lfloor p/q \rfloor} + \frac{2^p}{n} + \lambda^2\right)^{\frac{1+\gamma}{2+\gamma}}.
\end{aligned}$$

Here $C_\gamma < \infty$ depends only on $(\kappa, \beta, d, \tau, L, q, \Lambda, \gamma, C_{\mathrm{margin}})$ and the constants implicit in $|I| \asymp n$. This proves the excess risk bound in Theorem 3.11.

Choose $p = \lfloor q \log n / [q \log 2 - 2s \log(1-\alpha)] \rfloor$ and $\lambda \asymp \sqrt{2^p/n}$. By the choice of $p$,

$$(1-\alpha)^{2s\lfloor p/q \rfloor} + \frac{2^p}{n} + \lambda^2 \lesssim n^{\frac{2s\log(1-\alpha)}{q\log 2 - 2s\log(1-\alpha)}}.$$

Substituting into the excess risk bound just established gives

$$\mathbb{E}[R(\widehat{g}_M) - R(g^*)] \lesssim n^{\frac{2s(1+\gamma)\log(1-\alpha)}{(2+\gamma)\{q\log 2 - 2s\log(1-\alpha)\}}}.$$

For midpoint splits $\alpha = 1/2$ with $q = d$, the rate becomes $n^{-\frac{2s(1+\gamma)}{(2s+d)(2+\gamma)}}$. □

### A.5 Extension to Regression

We assume $P_{\mathbf{X}} = \mathrm{Unif}([0,1]^d)$ and consider i.i.d. observations $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ from the model $Y = f(\mathbf{X}) + \xi$, where $\mathbb{E}[\xi \mid \mathbf{X}] = 0$ and $\mathbb{E}[\xi^2 \mid \mathbf{X}] \le \sigma^2$ almost surely for some finite $\sigma^2 > 0$. Suppose that $f \in \mathcal{C}^{\kappa,\beta}$ as in Definition 3.6, with $\kappa \in \mathbb{N}_0$, $\beta \in (0,1]$, radius $L > 0$, and smoothness order $s = \kappa + \beta$.

The regression estimator uses the cyclic partition, subsampling scheme, polynomial basis, and penalty matrix from Section 2, with the same notation. The tree randomization $\Theta$ is independent of $\mathcal{D}_n$, and $|I| \asymp n$. For each leaf $A \in \Pi_p$, define the population coefficient vector by $\boldsymbol{\omega}_A = \mathbb{E}[\boldsymbol{\phi}_A(\mathbf{X}) f(\mathbf{X}) \mid \mathbf{X} \in A]$. For $\lambda > 0$ and $|I_A| > 0$, define

$$\widetilde{\boldsymbol{\omega}}_A = \arg\min_{\boldsymbol{\omega} \in \mathbb{R}^D} \left\{ \frac{1}{|I_A|} \sum_{i \in I_A} \left(Y_i - \boldsymbol{\phi}_A(\mathbf{X}_i)^\top \boldsymbol{\omega}\right)^2 + \lambda \boldsymbol{\omega}^\top J \boldsymbol{\omega} \right\}. \tag{27}$$

Project this vector onto the Euclidean ball of radius $R_f = \max\{1, L\}$ by

$$\widehat{\boldsymbol{\omega}}_A = \frac{\widetilde{\boldsymbol{\omega}}_A}{\max\{1, \|\widetilde{\boldsymbol{\omega}}_A\|_2 / R_f\}}.$$

Orthonormality and Definition 3.6 imply $\|\boldsymbol{\omega}_A\|_2^2 \le \mathbb{E}[f(\mathbf{X})^2 \mid \mathbf{X} \in A] \le L^2$, so the population coefficient vector belongs to the projection ball. For $\mathbf{x} \in A$, set $\widehat{f}(\mathbf{x}, \Theta) = \boldsymbol{\phi}_A(\mathbf{x})^\top \widehat{\boldsymbol{\omega}}_A$ when $|I_A| > 0$, and $\widehat{f}(\mathbf{x}, \Theta) = |I|^{-1} \sum_{i \in I} Y_i$ when $|I_A| = 0$.

For an integer $M \ge 1$, draw i.i.d. copies $\Theta_1, \ldots, \Theta_M$ of $\Theta$, independently of $\mathcal{D}_n$, and define the regression forest by averaging:

$$\widehat{f}_M(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \widehat{f}(\mathbf{x}, \Theta_m).$$

**Lemma A.10.** *Let $\Pi_p = \Pi_p(\Theta)$ and define the population polynomial projection by $f_{\Pi_p}(\mathbf{x}, \Theta) = \sum_{A \in \Pi_p} \boldsymbol{\omega}_A^\top \boldsymbol{\phi}_A(\mathbf{x}) \mathbf{1}\{\mathbf{x} \in A\}$. For every realization of $\Theta$,*

$$\mathbb{E}\left[ \left\| \widehat{f}(\cdot, \Theta) - f_{\Pi_p}(\cdot, \Theta) \right\|_{L_2}^2 \,\middle|\, \Theta \right] \le C_{\mathrm{est,reg}} \left( \frac{2^p}{|I|} + \lambda^2 \right),$$

*where $C_{\mathrm{est,reg}} < \infty$ depends only on $(\kappa, d, L, \sigma^2)$ and the constants implicit in $|I| \asymp n$. Moreover, choosing $\lambda \asymp \sqrt{2^p/n}$ yields*

$$\mathbb{E}\left[ \left\| \widehat{f}(\cdot, \Theta) - f_{\Pi_p}(\cdot, \Theta) \right\|_{L_2}^2 \,\middle|\, \Theta \right] = \mathcal{O}(2^p/n).$$

**Theorem A.11.** *Assume that Assumption 3.1 holds. Suppose that $f \in \mathcal{C}^{\kappa,\beta}$. Then*

$$\mathbb{E}\left[ \|\widehat{f}_M - f\|_{L_2}^2 \right] \le C_{\mathrm{reg}} \left( (1-\alpha)^{2s\lfloor p/q \rfloor} + \frac{2^p}{n} + \lambda^2 \right),$$

*where $C_{\mathrm{reg}} < \infty$ depends only on $(\kappa, \beta, d, L, q, \Lambda, \sigma^2)$ and the constants implicit in $|I| \asymp n$. Moreover, choosing $p = \left\lfloor \frac{q \log n}{q \log 2 - 2s \log(1-\alpha)} \right\rfloor$ and $\lambda \asymp \sqrt{2^p/n}$ yields*

$$\mathbb{E}\left[ \|\widehat{f}_M - f\|_{L_2}^2 \right] \lesssim n^{\frac{2s \log(1-\alpha)}{q \log 2 - 2s \log(1-\alpha)}}.$$

For centered splits ($\alpha = 1/2$) along a basis ($q = d$), the stated tuning choices yield $\mathcal{O}(n^{-2s/(2s+d)})$ over the full domain $[0,1]^d$. This matches the classical minimax lower bound for integrated squared error over Hölder classes (Stone, 1982). Thus the regression forest is minimax-rate optimal.

*Proof of Lemma A.10.* Condition on $\Theta = (I, T)$ and write $\Pi_p = \Pi_p(\Theta)$. The partition and subsample index set are then deterministic, with $|\Pi_p| = 2^p$. Since $\Theta$ is independent of the training sample, the observations indexed by $I$ remain i.i.d. Unless stated otherwise, all expectations and probabilities below are conditional on $\Theta$. For $A \in \Pi_p$, write $p_A = P_{\mathbf{X}}(A) > 0$. Lemma A.2 gives

$$\sup_{A \in \Pi_p} \sup_{\mathbf{x} \in A} \|\boldsymbol{\phi}_A(\mathbf{x})\|_2^2 \le L_{\mathrm{lev}}, \tag{28}$$

where $L_{\mathrm{lev}}$ depends only on $(\kappa, d)$.

Orthonormality and Definition 3.6 imply

$$\|\boldsymbol{\omega}_A\|_2^2 = \mathbb{E}[f_{\Pi_p}(\mathbf{X}, \Theta)^2 \mid \mathbf{X} \in A] \le \mathbb{E}[f(\mathbf{X})^2 \mid \mathbf{X} \in A] \le L^2.$$

For an empty leaf, write the prescribed constant prediction in coefficient form as

$$\widehat{\boldsymbol{\omega}}_A = \left( \frac{1}{|I|} \sum_{i \in I} Y_i, 0, \ldots, 0 \right)^\top.$$

This representation follows from $\phi_{A,1} \equiv 1$. Since $\boldsymbol{\omega}_A$ belongs to the Euclidean ball of radius $R_f = \max\{1, L\}$, projection onto this ball gives

$$\|\widehat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2 \le \|\widetilde{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2 \qquad \text{when } |I_A| > 0.$$

Moreover, on every occupied leaf,

$$\|\widehat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2^2 \le (R_f + L)^2 \le 4R_f^2. \tag{29}$$

To analyze the empty-leaf prediction, condition on the training features $(\mathbf{X}_i)_{i \in I}$. Independence and the noise assumptions give

$$\mathbb{E}\left[ \frac{1}{|I|} \sum_{i \in I} Y_i \,\middle|\, (\mathbf{X}_i)_{i \in I} \right] = \frac{1}{|I|} \sum_{i \in I} f(\mathbf{X}_i),$$

$$\mathrm{Var}\left( \frac{1}{|I|} \sum_{i \in I} Y_i \,\middle|\, (\mathbf{X}_i)_{i \in I} \right) \le \frac{\sigma^2}{|I|}.$$

Since $\big||I|^{-1}\sum_{i\in I} f(\mathbf{X}_i)\big| \le L$ and $\|\boldsymbol{\omega}_A\|_2 \le L$, on $\{|I_A|=0\}$ we obtain

$$\begin{aligned}
&\mathbb{E}\big[\|\widehat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2^2 \,\big|\, (\mathbf{X}_i)_{i\in I}\big] \\
&\quad = \left\| \left( \frac{1}{|I|}\sum_{i\in I} f(\mathbf{X}_i), 0, \ldots, 0 \right)^\top - \boldsymbol{\omega}_A \right\|_2^2 \\
&\qquad + \mathrm{Var}\left( \frac{1}{|I|}\sum_{i\in I} Y_i \,\middle|\, (\mathbf{X}_i)_{i\in I} \right) \\
&\quad \le 4L^2 + \frac{\sigma^2}{|I|}.
\end{aligned}$$

The event $\{|I_A|=0\}$ is determined by these features. Taking expectations therefore yields

$$\mathbb{E}\big[\|\widehat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2^2 \mathbf{1}\{|I_A|=0\}\big] \le \left(4L^2 + \frac{\sigma^2}{|I|}\right)\mathbb{P}(|I_A|=0). \tag{30}$$

Use the empirical Gram matrix $\widehat{G}_A$ defined in Lemma A.3. For $|I_A|>0$, define

$$\boldsymbol{\Delta}_A = \frac{1}{|I_A|}\sum_{i\in I_A} \boldsymbol{\phi}_A(\mathbf{X}_i)\big(Y_i - f_{\Pi_p}(\mathbf{X}_i,\Theta)\big),$$

and set $\boldsymbol{\Delta}_A = \mathbf{0}$ when $|I_A|=0$. Conditional on a nonzero leaf count, the observations in $A$, ordered by their original sample indices, are i.i.d. from $P(\cdot \mid \mathbf{X}\in A)$. The conditional mean-zero noise assumption and orthogonality of the population projection give

$$\mathbb{E}\big[\boldsymbol{\phi}_A(\mathbf{X})\big(Y - f_{\Pi_p}(\mathbf{X},\Theta)\big) \,\big|\, \mathbf{X}\in A\big] = 0.$$

Also,

$$\begin{aligned}
&\mathbb{E}\Big[\big(Y - f_{\Pi_p}(\mathbf{X},\Theta)\big)^2 \,\Big|\, \mathbf{X}\in A\Big] \\
&\quad = \mathbb{E}[\xi^2 \mid \mathbf{X}\in A] + \mathbb{E}\Big[\big(f(\mathbf{X}) - f_{\Pi_p}(\mathbf{X},\Theta)\big)^2 \,\Big|\, \mathbf{X}\in A\Big] \\
&\quad \le \sigma^2 + L^2,
\end{aligned}$$

where the last inequality uses the admissibility of the zero polynomial as an approximation to $f$. Conditional independence and (28) therefore yield, on $\{|I_A|>0\}$,

$$\mathbb{E}\big[\|\boldsymbol{\Delta}_A\|_2^2 \,\big|\, |I_A|\big] \le \frac{L_{\mathrm{lev}}(\sigma^2+L^2)}{|I_A|}. \tag{31}$$

The positive-definiteness argument in Lemma A.3 depends only on the design matrix and $\lambda>0$, and hence applies to the regression objective. On every occupied leaf, the normal equations for (27) give

$$\widetilde{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A = (\widehat{G}_A + \lambda J)^{-1}(\boldsymbol{\Delta}_A - \lambda J \boldsymbol{\omega}_A). \tag{32}$$

Use the event $\mathcal{E}_A(1/2)$ defined in Lemma A.6. On this event, the inverse in (32) has operator norm at most 2. Since $\|J\boldsymbol{\omega}_A\|_2 \le L$, it follows that

$$\begin{aligned}
&\|\widehat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2^2 \mathbf{1}\{\mathcal{E}_A(1/2)\} \\
&\quad \le 8\|\boldsymbol{\Delta}_A\|_2^2 \mathbf{1}\{\mathcal{E}_A(1/2)\} + 8L^2\lambda^2 \mathbf{1}\{\mathcal{E}_A(1/2)\}.
\end{aligned}$$

Taking expectations and dropping the indicator in the nonnegative residual term, we use (31), (29) on occupied leaves, and (30) on empty leaves. Since $L \le R_f$ and $\{|I_A|=0\} \subseteq \mathcal{E}_A(1/2)^c$, we obtain

$$\begin{aligned}
\mathbb{E}\|\widehat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2^2 \le 8L_{\mathrm{lev}}(\sigma^2+L^2)\mathbb{E}\left[\frac{\mathbf{1}\{|I_A|>0\}}{|I_A|}\right] + 8L^2\lambda^2 \\
+ 4R_f^2\mathbb{P}(\mathcal{E}_A(1/2)^c) + \frac{\sigma^2}{|I|}\mathbb{P}(|I_A|=0).
\end{aligned} \tag{33}$$

The fraction is defined to be zero when $|I_A|=0$.

Because $|I_A| \sim \mathrm{Bin}(|I|, p_A)$, Lemma A.1 implies

$$p_A \mathbb{E}\left[\frac{\mathbf{1}\{|I_A| > 0\}}{|I_A|}\right] \le \frac{2}{|I|+1} \le \frac{2}{|I|}.$$

Also,

$$p_A \mathbb{P}(|I_A| = 0) = p_A(1-p_A)^{|I|} \le p_A e^{-|I|p_A} \le \frac{1}{e|I|}.$$

Hence the additional empty-leaf term in (33), after multiplication by $p_A$, is at most $\sigma^2/(e|I|^2)$, and thus at most $\sigma^2/(e|I|)$ because $|I| \ge 1$. The Gram matrix depends only on the features, so Lemma A.6 also applies and gives $p_A \mathbb{P}(\mathcal{E}_A(1/2)^c) \lesssim 1/|I|$, with an implicit constant depending only on $(\kappa, d)$. Consequently, (33) yields

$$p_A \mathbb{E}\|\widehat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2^2 \le C_{\mathrm{est,reg}} \left(\frac{1}{|I|} + \lambda^2 p_A\right), \tag{34}$$

where $C_{\mathrm{est,reg}} < \infty$ depends only on $(\kappa, d, L, \sigma^2)$ and the constants implicit in $|I| \asymp n$.

Orthonormality on each leaf gives

$$\|\widehat{f}(\cdot,\Theta) - f_{\Pi_p}(\cdot,\Theta)\|_{L_2}^2 = \sum_{A \in \Pi_p} p_A \|\widehat{\boldsymbol{\omega}}_A - \boldsymbol{\omega}_A\|_2^2.$$

Summing (34) over the $2^p$ leaves, using $\sum_{A\in\Pi_p} p_A = 1$, and restoring the conditioning on $\Theta$ yields

$$\mathbb{E}\left[\|\widehat{f}(\cdot,\Theta) - f_{\Pi_p}(\cdot,\Theta)\|_{L_2}^2 \,\middle|\, \Theta\right] \le C_{\mathrm{est,reg}} \left(\frac{2^p}{|I|} + \lambda^2\right).$$

Since $|I| \asymp n$, choosing $\lambda \asymp \sqrt{2^p/n}$ gives the stated $\mathcal{O}(2^p/n)$ bound. □

*Proof of Theorem A.11.* Condition on $\Theta = (I, T)$ and write $\Pi_p = \Pi_p(\Theta)$. The function $f_{\Pi_p}(\cdot,\Theta)$ is the $L_2$-orthogonal projection of $f$ onto the piecewise degree-$\kappa$ polynomial space. Every realized tree estimate belongs to this space, including on empty leaves. The Pythagorean identity gives

$$\|\widehat{f}(\cdot,\Theta) - f\|_{L_2}^2 = \|f_{\Pi_p}(\cdot,\Theta) - f\|_{L_2}^2 + \|\widehat{f}(\cdot,\Theta) - f_{\Pi_p}(\cdot,\Theta)\|_{L_2}^2. \tag{35}$$

The scalar Taylor approximation argument in the proof of Lemma A.7, applied to $f$, gives a finite $K_0 = K_0(\kappa, \beta, d)$ such that

$$\mathbb{E}\left[\left(f(\mathbf{X}) - f_{\Pi_p}(\mathbf{X},\Theta)\right)^2 \,\middle|\, \mathbf{X} \in A\right] \le K_0^2 L^2 \,\mathrm{diam}(A)^{2s}.$$

Thus Lemma 3.4 implies

$$\|f - f_{\Pi_p}(\cdot,\Theta)\|_{L_2}^2 \le K_0^2 L^2 (\Lambda q d)^s (1-\alpha)^{2s\lfloor p/q \rfloor}. \tag{36}$$

Combining (35) and (36) with Lemma A.10 and $|I| \asymp n$ yields

$$\mathbb{E}\left[\|\widehat{f}(\cdot,\Theta) - f\|_{L_2}^2 \,\middle|\, \Theta\right] \le C_{\mathrm{reg}} \left((1-\alpha)^{2s\lfloor p/q \rfloor} + \frac{2^p}{n} + \lambda^2\right), \tag{37}$$

where $C_{\mathrm{reg}} < \infty$ depends only on $(\kappa, \beta, d, L, q, \Lambda, \sigma^2)$ and the constants implicit in $|I| \asymp n$. The bound is uniform over $\Theta$ and therefore also holds after averaging over the tree randomization.

For every realization of the training sample and the $M$ tree randomizations, convexity of the squared $L_2$ norm gives

$$\|\widehat{f}_M - f\|_{L_2}^2 \le \frac{1}{M} \sum_{m=1}^{M} \|\widehat{f}(\cdot,\Theta_m) - f\|_{L_2}^2.$$

Taking expectations and applying (37) proves the risk bound in Theorem A.11.

Choose $p = \lfloor q \log n / [q \log 2 - 2s \log(1-\alpha)] \rfloor$ and $\lambda \asymp \sqrt{2^p/n}$. By the choice of $p$,

$$\frac{2^p}{n} \le n^{\frac{2s\log(1-\alpha)}{q\log 2 - 2s\log(1-\alpha)}},$$

$$(1-\alpha)^{2s\lfloor p/q \rfloor} \le (1-\alpha)^{-2s} n^{\frac{2s\log(1-\alpha)}{q\log 2 - 2s\log(1-\alpha)}}.$$

Substituting these bounds and $\lambda^2 \asymp 2^p/n$ into the risk bound just established gives

$$\mathbb{E}\left[\|\widehat{f}_M - f\|_{L_2}^2\right] \lesssim n^{\frac{2s\log(1-\alpha)}{q\log 2 - 2s\log(1-\alpha)}}.$$

For centered splits $\alpha = 1/2$ with $q = d$, the exponent is $-2s/(2s+d)$, which gives the stated midpoint rate. □

### A.6 Proofs of the Auxiliary Lemmas

Write $p_A = P_{\mathbf{X}}(A) = \int_A d\mathbf{x}$. Conditional on any realization of $\Theta = (I, T)$, $|I_A| \sim \mathrm{Bin}(|I|, p_A)$, since the tree randomization is independent of the training sample. Throughout this subsection, expectations are conditional on $\Theta$, with this conditioning suppressed in the notation.

*Proof of Lemma A.1.* Since $1/k \le 2/(k+1)$ for every integer $k \ge 1$, $\mathbb{E}[\mathbf{1}\{|I_A| > 0\}/|I_A|] \le 2\mathbb{E}[1/(|I_A|+1)]$. The binomial formula gives

$$\mathbb{E}\left[\frac{1}{|I_A|+1}\right] = \int_0^1 (1 - p_A + p_A t)^{|I|}\, dt = \frac{1-(1-p_A)^{|I|+1}}{(|I|+1)p_A} \le \frac{1}{(|I|+1)p_A},$$

which proves (7). For $a > 0$, the same formula yields

$$\begin{aligned} p_A \mathbb{E}\left[e^{-a|I_A|}\right] &= p_A\big(1 - (1-e^{-a})p_A\big)^{|I|} \\ &\le p_A \exp\big(-|I|(1-e^{-a})p_A\big) \le \frac{1}{e|I|(1-e^{-a})}. \end{aligned}$$

The last inequality follows from $ue^{-u} \le e^{-1}$ for $u \ge 0$ and proves (8). □

*Proof of Lemma A.2.* Every leaf is a convex polytope up to its boundary convention. Since $\alpha \in (0, 1/2]$, each cut lies strictly inside the projection interval of its parent. Both children of a cell with nonempty interior therefore have nonempty interior. Starting from $[0,1]^d$, induction gives $p_A > 0$ for every leaf. Replacing $A$ by its closure preserves its integrals and polynomial suprema, so we may regard $A$ as a compact convex body.

For $\kappa = 0$, the claim holds with $L_{\mathrm{lev}} = 1$. For $\kappa \ge 1$, an affine rescaling to unit volume and Carbery & Wright (2001, Theorem 1) give a finite $L_{\mathrm{lev}}$, depending only on $(\kappa, d)$, such that, for every polynomial $r \in \mathcal{P}_\kappa(A)$,

$$\sup_{\mathbf{x}\in A} |r(\mathbf{x})|^2 \le \frac{L_{\mathrm{lev}}}{p_A} \int_A |r(\mathbf{x})|^2\, d\mathbf{x}. \tag{38}$$

For every $\mathbf{u} \in \mathbb{R}^D$ with $\|\mathbf{u}\|_2 = 1$, orthonormality gives $p_A^{-1}\int_A |\mathbf{u}^\top \boldsymbol{\phi}_A(\mathbf{x})|^2\, d\mathbf{x} = 1$. Applying (38) to $\mathbf{u}^\top \boldsymbol{\phi}_A$ and using $\sup_{\|\mathbf{u}\|_2=1} |\mathbf{u}^\top \boldsymbol{\phi}_A(\mathbf{x})|^2 = \|\boldsymbol{\phi}_A(\mathbf{x})\|_2^2$ proves (9). □

*Proof of Lemma A.3.* For every $\mathbf{v} = (v_1, \dots, v_D)^\top \in \mathbb{R}^D$,

$$\mathbf{v}^\top\big(\widehat{G}_A + \lambda J\big)\mathbf{v} = \frac{1}{|I_A|}\sum_{i\in I_A}\big(\mathbf{v}^\top \boldsymbol{\phi}_A(\mathbf{X}_i)\big)^2 + \lambda \sum_{j=2}^{D} v_j^2.$$

Both terms are nonnegative. If their sum is zero, then $v_j = 0$ for every $j \ge 2$. Since $\phi_{A,1} \equiv 1$, the first term then equals $v_1^2$, so $\mathbf{v} = \mathbf{0}$. Thus $\widehat{G}_A + \lambda J$ is positive definite. The ridge objective is strictly convex and coercive, and its first-order condition gives (10). □

*Proof of Lemma A.4.* By orthonormality and contraction of the orthogonal projection onto $\mathcal{P}_\kappa(A)$,

$$\|\Omega_A\|_F^2 = \frac{1}{p_A}\int_A \|\boldsymbol{\eta}_{\Pi_p}(\mathbf{x}, \Theta)\|_2^2\, d\mathbf{x} \le \frac{1}{p_A}\int_A \|\boldsymbol{\eta}(\mathbf{x})\|_2^2\, d\mathbf{x} \le 1,$$

since $\boldsymbol{\eta}(\mathbf{x})$ is a probability vector.

For $|I_A| > 0$, (5) defines $\widehat{\Omega}_A$ as the Euclidean projection of $\widetilde{\Omega}_A$ onto the unit Frobenius ball. Since $\Omega_A$ belongs to this ball, nonexpansiveness gives $\|\widehat{\Omega}_A - \Omega_A\|_F \le \|\widetilde{\Omega}_A - \Omega_A\|_F$. Together with orthonormality, this proves (11).

Moreover, $\|\widehat{\Omega}_A\|_F \le 1$, so $p_A^{-1}\int_A \|\widehat{\boldsymbol{\eta}}(\mathbf{x}, \Theta)\|_2^2\, d\mathbf{x} \le 1$. For an empty leaf, the score is a constant probability vector, and the same bound holds. The corresponding bound for $\boldsymbol{\eta}_{\Pi_p}(\cdot, \Theta)$ follows from $\|\Omega_A\|_F \le 1$. The triangle inequality bounds their normalized $L_2$ distance by 2. Squaring proves (12). □

*Proof of Lemma A.5.* Conditional on $\Theta$ and $|I_A| = k$, list the feature vectors in $A$ as $\mathbf{X}_{A,1}, \dots, \mathbf{X}_{A,k}$ in increasing order of their original sample indices. These vectors are i.i.d. from $P_{\mathbf{X}}(\cdot \mid \mathbf{X} \in A) = \mathrm{Unif}(A)$. For $r = 1, \dots, k$, define $Z_r = \boldsymbol{\phi}_A(\mathbf{X}_{A,r})\boldsymbol{\phi}_A(\mathbf{X}_{A,r})^\top$. These matrices are conditionally independent and positive semidefinite. Orthonormality gives $\mathbb{E}[Z_r \mid \Theta, |I_A| = k] = \mathrm{diag}(1, \dots, 1)$, and Lemma A.2 gives $\|Z_r\|_{\mathrm{op}} = \|\boldsymbol{\phi}_A(\mathbf{X}_{A,r})\|_2^2 \le L_{\mathrm{lev}}$.

Since $\widehat{G}_A = k^{-1}\sum_{r=1}^k Z_r$, the matrix Chernoff lower-tail inequality yields

$$\begin{aligned}\mathbb{P}(\mathcal{E}_A(\zeta)^c \,|\, \Theta, |I_A| = k) &= \mathbb{P}\Big(\lambda_{\min}(\widehat{G}_A) < \zeta \,\Big|\, \Theta, |I_A| = k\Big) \\ &\le D\left(\frac{e^{\zeta-1}}{\zeta^\zeta}\right)^{k/L_{\mathrm{lev}}} = De^{-a_\zeta k}.\end{aligned}$$

The coefficient $a_\zeta$ is positive because $1 - \zeta + \zeta\log\zeta > 0$ for $\zeta \in (0,1)$. □

*Proof of Lemma A.6.* For every $k \in \{0, \dots, |I|\}$ with $\mathbb{P}(|I_A| = k \mid \Theta) > 0$,

$$\mathbb{P}(\mathcal{E}_A(\zeta)^c \,|\, \Theta, |I_A| = k) \le De^{-a_\zeta k}.$$

For $k \ge 1$, this follows from Lemma A.5; for $k = 0$, the left-hand side equals $1 \le D$. Averaging over $|I_A|$ and applying (8) gives

$$p_A\mathbb{P}(\mathcal{E}_A(\zeta)^c \mid \Theta) \le Dp_A\mathbb{E}\Big[e^{-a_\zeta|I_A|} \mid \Theta\Big] \le \frac{D}{e|I|(1-e^{-a_\zeta})}.$$

□

*Proof of Lemma A.7.* For each $A \in \Pi$, choose $\mathbf{x}_A \in A$ and let $\mathbf{T}_A$ be the vector of degree-$\kappa$ Taylor polynomials of $\eta_1, \dots, \eta_\tau$ at $\mathbf{x}_A$. The line segment joining $\mathbf{x}_A$ to any $\mathbf{x} \in A$ lies in $[0,1]^d$. Hölder smoothness and Taylor's theorem therefore give

$$\sup_{\mathbf{x}\in A}\|\boldsymbol{\eta}(\mathbf{x}) - \mathbf{T}_A(\mathbf{x})\|_2^2 \le K_{\mathrm{app}}L^2\,\mathrm{diam}(A)^{2s},$$

for a finite $K_{\mathrm{app}}$ depending only on $(\kappa, \beta, d, \tau)$. On each cell, $\boldsymbol{\eta}_\Pi$ is the orthogonal projection of $\boldsymbol{\eta}$ onto $\mathcal{P}_\kappa(A)^\tau$, and $\mathbf{T}_A \in \mathcal{P}_\kappa(A)^\tau$. Hence

$$\begin{aligned}\|\boldsymbol{\eta} - \boldsymbol{\eta}_\Pi\|_{L_2}^2 &\le \sum_{A\in\Pi}\int_A \|\boldsymbol{\eta}(\mathbf{x}) - \mathbf{T}_A(\mathbf{x})\|_2^2\,d\mathbf{x} \\ &\le K_{\mathrm{app}}L^2\sum_{A\in\Pi} p_A\,\mathrm{diam}(A)^{2s}.\end{aligned}$$

Since $\sum_{A\in\Pi} p_A = 1$, (15) follows. □

*Proof of Lemma A.8.* By the definition of $\widetilde{g}$, for every $\mathbf{x}$,

$$\begin{aligned}\eta_{g^*(\mathbf{x})}(\mathbf{x}) - \eta_{\widetilde{g}(\mathbf{x})}(\mathbf{x}) &\le \eta_{g^*(\mathbf{x})}(\mathbf{x}) - \widetilde{\eta}_{g^*(\mathbf{x})}(\mathbf{x}) \\ &\quad + \widetilde{\eta}_{\widetilde{g}(\mathbf{x})}(\mathbf{x}) - \eta_{\widetilde{g}(\mathbf{x})}(\mathbf{x}) \\ &\le 2\max_{1\le c\le\tau}|\widetilde{\eta}_c(\mathbf{x}) - \eta_c(\mathbf{x})|.\end{aligned}$$

For every classifier $g$, $R(g) = \int_{[0,1]^d}[1 - \eta_{g(\mathbf{x})}(\mathbf{x})]\,dP_{\mathbf{X}}(\mathbf{x})$. Integrating the pointwise bound and bounding the largest absolute component by the Euclidean norm proves (16). Under the uniform design, Cauchy–Schwarz gives $\int_{[0,1]^d}\|\widetilde{\boldsymbol{\eta}}(\mathbf{x}) - \boldsymbol{\eta}(\mathbf{x})\|_2\,d\mathbf{x} \le \|\widetilde{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_{L_2}$, which proves (17). □

*Proof of Lemma A.9.* Since $w_j(K) > 0$ and $\alpha \in (0, 1/2]$, $\inf_{\mathbf{x}\in K}\mathbf{b}_j^\top\mathbf{x} < m_j(K) < \sup_{\mathbf{x}\in K}\mathbf{b}_j^\top\mathbf{x}$. Thus both children are nonempty. Their projection intervals give

$$\begin{aligned}w_j(K^{\mathrm{L}}) &\le m_j(K) - \inf_{\mathbf{x}\in K}\mathbf{b}_j^\top\mathbf{x} = \alpha w_j(K), \\ w_j(K^{\mathrm{R}}) &\le \sup_{\mathbf{x}\in K}\mathbf{b}_j^\top\mathbf{x} - m_j(K) = (1-\alpha)w_j(K).\end{aligned}$$

For $K' \subseteq K$, the supremum of each projection cannot increase and its infimum cannot decrease, proving the last claim. □