Fréchet Regression with Mondrian Forests: Finite-Sample Guarantees and Ensemble Benefits

Rui Qiu, Fang Yao, Zhou Yu

***Abstract*—Fréchet random forests extend the power of classical random forests to general metric spaces, offering promising advantages over traditional methods, especially in high-dimensional settings. While forests have empirically outperformed individual trees in Fréchet regression, the theoretical basis for this improvement remains largely unexplored. This paper fills this gap by establishing non-asymptotic upper bounds for the prediction risk of Fréchet Mondrian forests, complementing the existing literature that primarily focuses on asymptotic analysis. We demonstrate that, under suitable regularity conditions, Fréchet Mondrian forests attain convergence rates comparable to their Euclidean counterparts. Moreover, under higher-order smoothness assumptions and with a sufficient number of trees, Fréchet forests achieve faster convergence than individual Fréchet trees, thereby providing a rigorous theoretical justification for the benefit of ensembles in non-Euclidean regression problems. The effectiveness of the proposed method is further corroborated through simulation studies across diverse settings, including probability distributions, symmetric positive-definite matrices, and spherical data.**

***Index Terms*—Ensemble learning, Fréchet regression, metric space, Mondrian forest, non-asymptotic analysis.**

I. Introduction

NON-EUCLIDEAN data analysis has gained significant attention recently due to the increasing availability of structured data in fields like medical imaging, computational biology, and network science. As a prominent line of research, Fréchet regression extends traditional regression models, where responses are scalar or vector-valued, to accommodate data such as probability distributions, shapes, or networks, which naturally lie in nonlinear spaces (e.g., Wasserstein or manifold spaces). The methodology builds on Fréchet's pioneering concept of the intrinsic mean [1] in metric spaces, defined as the point minimizing the expected squared distance. Seminal works by [2] introduced a conditional Fréchet mean framework and developed global and local linear methods to model relationships between scalar predictors and non-Euclidean responses, leveraging the intrinsic geometry of the response space. Subsequent advancements, including extensions to longitudinal data [3], [4], single-index models [5], [6], nonlinear global regression [7] and regularization techniques [8], have further expanded the scope and applicability of Fréchet regression, solidifying its role as a critical tool for modern statistical analysis. While the aforementioned works focus on the asymptotic theory of Fréchet regression methods, [9] shifted attention to its non-asymptotic properties and revealed important insights into finite-sample behavior.

Recently, research on Fréchet regression has increasingly shifted toward ensemble models. Random forests, introduced by [10], are among the most widely used ensemble learning methods, combining predictions from multiple decision trees to improve accuracy and mitigate overfitting. Random forests are appealing due to their few tuning parameters, parallel training, robustness to feature scaling and noise, reasonable computational cost, and suitability for high-dimensional settings. Over the years, extensive theoretical studies have established the consistency and asymptotic properties [11]–[18] of random forests, while practical enhancements have broadened their applicability across various domains. Among these advancements, Mondrian forests, proposed by [19], stand out as an important development due to the favorable geometric properties of Mondrian splitting. [20] introduced a Mondrian forest algorithm for online learning. In offline learning [21], Mondrian forests have been demonstrated as the first class of random forests to achieve the minimax optimal rate in arbitrary dimensions under nonparametric assumptions [22]. More importantly, Mondrian forests exhibit faster convergence rates than individual trees in higher-order Hölder continuous regression, thereby highlighting the role of exogenous randomness in the model aggregation algorithm. Later, [23] provided the asymptotic normal distribution of Mondrian forests and developed valid statistical inference with a debiasing approach. More recently, [24] introduced oblique variants of Mondrian forests, called Tessellation forests, which also achieve minimax optimal rates in arbitrary dimensions with general split directions.

The aforementioned studies focus on regression models with scalar-valued responses. For non-Euclidean responses, local linear Fréchet regression [2], which relies on nonparametric kernel smoothing without model structure assumptions, faces challenges in high-dimensional settings [25], [26]. This limi-

The work of Fang Yao was supported in part by the National Natural Science Foundation of China under Grant 12292981 and 12288101, in part by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM118, in part by the New Cornerstone Science Foundation through the Xplorer Prize, and in part by the LMAM and the Fundamental Research Funds for the Central Universities, Peking University (LMEQF). The work of Rui Qiu was supported in part by the National Natural Science Foundation of China under Grant 12501346, and in part by the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant BX20250069 and 2024M760060. The work of Zhou Yu was supported in part by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM904, in part by the National Natural Science Foundation of China under Grant 12371289, and in part by the Shanghai Pilot Program for Basic Research under Grant TQ20220105. (Corresponding author: Fang Yao.)

Rui Qiu and Fang Yao are with the School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing 100871, China (e-mail: rqiu_stat@outlook.com; fyao@math.pku.edu.cn).

Zhou Yu is with the School of Statistics, KLATASDS-MOE, East China Normal University, Shanghai 200241, China (e-mail: zyu@stat.ecnu.edu.cn).

tation hinders the broader applicability of Fréchet regression in real-world applications. To this end, extending random forests to Fréchet regression is a promising direction. When the predictor dimension is relatively high, these methods are expected to outperform existing Fréchet regression approaches, given the favorable properties of random forests. [27] proposed the concept of Fréchet trees and Fréchet random forests. The Fréchet mean of the tree outputs serves as a prediction in Fréchet random forests, following the tree-averaging strategy of Euclidean random forests. In contrast, [28] followed the framework of Fréchet regression [2]. They viewed random forests as powerful local weight generators and applied the generated weights to local Fréchet regression. Although these are two distinct methods, both treat Euclidean random forests as special cases. Both Fréchet random forest methods exhibit desirable asymptotic properties and demonstrate outstanding numerical performance in complex scenarios.

The integration of random forests into Fréchet regression provides a powerful tool for non-Euclidean data analysis. However, similar to the Euclidean case, the mechanisms underlying the effectiveness of Fréchet random forests are not yet fully understood, and this paper aims to bridge this gap. Due to the lack of linear structure, the analysis of Fréchet random forests must often proceed without explicit expressions, which creates substantial challenges for theory. Notably, the Euclidean random forest is a special case of the Fréchet random forest. A natural and important question is which properties of Euclidean random forests can be transferred to non-Euclidean settings. This paper has two primary objectives:

(1) To establish a non-asymptotic upper bound for the prediction risk of Fréchet random forests. The existing literature on Fréchet random forests has primarily focused on asymptotic convergence properties, with much of the Fréchet regression literature also centered on large sample theory. Establishing non-asymptotic convergence rates for Fréchet random forests will offer additional insights into the convergence behavior of these methods and provide theoretical support for their numerical performance in finite sample settings.

(2) To explore when Fréchet random forests outperform individual trees. In the Euclidean case, numerous studies have investigated the reasons behind the superior performance of random forests compared to individual trees, from both theoretical and empirical perspectives [21], [29]–[32]. However, it remains unclear whether this advantage extends to the non-Euclidean case. While Jensen's inequality ensures that Euclidean random forests perform at least as well as individual trees in terms of prediction risk, it does not apply in non-Euclidean spaces. Additionally, [27] only examined the convergence of individual trees, while [28] only studied random forests with infinite trees. Therefore, a comparison between Fréchet random forests and individual Fréchet trees is still an open question that warrants further exploration.

This paper adopts the Fréchet random forest strategy proposed by [28] and combines it with the Mondrian splitting criterion [21], resulting in a method we term the Fréchet Mondrian forest, which serves as the foundation for achieving the two objectives outlined above. Specifically, under suitable non-Euclidean assumptions and smoothness conditions, we establish the consistency and convergence rate of Fréchet Mondrian forests in arbitrary dimensions. Our results demonstrate that when the Mondrian parameter λ and the number M of trees in Fréchet random forests are properly tuned, we can achieve the same convergence rate as in the Euclidean case. To the best of our knowledge, this is the first non-asymptotic result for Fréchet random forests. Furthermore, under higher-order smoothness assumptions, we find that when the number of trees exceeds a certain threshold, the Fréchet Mondrian forest achieves faster convergence than an individual Fréchet Mondrian tree due to its smaller bias, which again aligns with the Euclidean case. These analyses highlight the importance of ensemble learning in Fréchet regression and offer valuable insights into the relationship between Fréchet regression and Euclidean regression.

The rest of the paper is organized as follows. Section II presents the necessary background of Fréchet regression and motivates the choice of our forest model. Section III develops the theoretical properties of the Fréchet Mondrian tree and forest, establishing their consistency and convergence rates under β-Hölder ($\beta \leq 1$) smoothness conditions. Section IV then focuses on the advantages of forests over individual trees under higher-order smoothness assumptions. Section V empirically investigates the predictive performance of Fréchet Mondrian forests across different response types. Finally, in Section VI, we provide a brief summary and discussion of our findings. All proofs are provided in the Appendices.

Some notation: We use c to represent general absolute constants whose value may change from line to line. If the value depends on some variables, we indicate them by an index. $\lfloor a \rfloor$ is the integer part of the real number a. $\|x\|_2$ is the l_2 norm of the vector $x = \left(x_{(1)}, \ldots, x_{(d)}\right)^\top$ and $\|A\|_{\mathrm{F}}$ is the Frobenius norm of the matrix $A = (a_{ij})$. Let $\mathcal{N}_0$ denote the set of non-negative integers and $\mathcal{R}_+^k$ be the subset of $\mathcal{R}^k$ consisting of points whose components are all positive. $\mathrm{B}(x, d, \varepsilon)$ denotes the ball of center x and radius ε with respect to the metric d. The d-diameter of a set A is defined by $\operatorname{diam}(A, d) = \sup_{x_1, x_2 \in A} d(x_1, x_2)$. $U([0,1]^p)$ denotes the uniform distribution over the p-dimensional unit cube $[0,1]^p$. $\mathcal{N}(a, b)$ denotes the normal distribution on $\mathcal{R}$ with mean a and variance b.

II. Motivation and Method

In this section, we introduce the background of the problem and the Fréchet random forest model used for the subsequent theoretical analysis.

A. Fréchet regression

Let (Ω, d) denote a metric space equipped with a specific metric d, and let $\mathcal{R}^p$ represent the p-dimensional Euclidean space. Given a probability space $(\mathcal{T}, \mathcal{B}_\mathcal{T}, P_\mathcal{T})$, where $\mathcal{B}_\mathcal{T}$ is the Borel σ-algebra in the domain $\mathcal{T}$ and $P_\mathcal{T}$ is a probability measure, we consider a random pair $Z = (X, Y) \in [0,1]^p \times \Omega$. As a measurable mapping from $\mathcal{T}$ to $[0,1]^p \times \Omega$, the joint law

of (X, Y) is represented by ν, such that $\nu(A) = P_{\mathcal{T}}(\tau \in \mathcal{T} : Z(\tau) \in A)$ for any Borel measurable set $A \subset [0,1]^p \times \Omega$. Let ν_X denote the marginal distribution of X. The conditional probability measure of Y given $X = x$ is assumed to exist.

Now we focus on the regression problem. In the special case where $\Omega = \mathcal{R}$, the objective of classical Euclidean regression is to estimate $m(x) = \mathbb{E}(Y|X = x)$. However, when the response Y resides on a general metric space without a linear structure, a similar definition of Fréchet regression becomes impractical, hindering the definition of conditional Fréchet mean [2]

$$\begin{aligned} m_{\oplus}(x) &= \operatorname*{argmin}_{\omega \in \Omega} F_x(\omega) \\ &:= \operatorname*{argmin}_{\omega \in \Omega} \mathbb{E}\left\{d^2(Y, \omega) \mid X = x\right\}. \end{aligned} \tag{1}$$

By replacing the metric d with the Euclidean distance when $\Omega = \mathcal{R}$, (1) returns to the classical Euclidean regression function $m(x)$.

B. Fréchet Mondrian tree

A Fréchet tree, referring to a tree that handles metric space-valued responses, splits the input space recursively from the root node (the entire input space). Each split divides a parent node into two child nodes along a chosen feature direction and cutoff point, determined by a specific criterion. After multiple splits, sufficiently small child nodes form leaf nodes, where sample points are used to estimate the conditional Fréchet mean. Tree variants differ primarily in their splitting criteria. This paper focuses on the Fréchet Mondrian tree, constructed via the Mondrian process (see [21]). A Mondrian partition of feature space can be sampled from the Mondrian process distribution $\mathrm{MP}(\lambda, C)$ using the recursive procedure $\mathrm{SampleMondrian}(C, \tau = 0, \lambda)$ described in Algorithm 1. The lifetime parameter λ governs partition complexity, with larger values resulting in deeper trees. Any sample drawn from $\mathrm{MP}(\lambda, [0,1]^p)$ generates a recursive partition of $[0,1]^p$, thereby resulting in a Fréchet Mondrian tree.

Algorithm 1 $\mathrm{SampleMondrian}(C, \tau, \lambda)$: Samples a Mondrian partition of C starting at time τ and continuing until time λ.

Inputs: A cell $C = \prod_{1 \le j \le p}[a_j, b_j]$, starting time τ, and lifetime parameter λ.
Sample a random variable $E_C \sim \mathrm{Exp}(|C|)$ with $|C| = \sum_{j=1}^{p}(b_j - a_j)$.
if $\tau + E_C \le \lambda$ **then**
  Sample a split dimension $J \in \{1, \ldots, p\}$ with probability $\mathbb{P}(J = j) = (b_j - a_j)/|C|$;
  Sample a split threshold S_J uniformly in $[a_J, b_J]$;
  Split C along the split (J, S_J): let $C_L = \left\{x \in C : x_{(J)} \le S_J\right\}$ and $C_R = C \setminus C_L$;
  return $\mathrm{SampleMondrian}(C_L, \tau + E_C, \lambda) \cup \mathrm{SampleMondrian}(C_R, \tau + E_C, \lambda)$.
else
  return $\{C\}$ (i.e., do not split C).
end if

We sample a partition Π_λ from $\mathrm{MP}(\lambda, [0,1]^p)$ to construct a Fréchet Mondrian tree T. Given training data $\{(X_i, Y_i)\}_{i=1}^n$ from ν, then $m_{\oplus}(x)$ can be predicted by the sample Fréchet mean of responses of the observations falling in the same leaf as x, that is,

$$\begin{aligned} \hat{m}_{\oplus}(x) &= \operatorname*{argmin}_{\omega \in \Omega} \hat{F}_x(\omega) \\ &:= \operatorname*{argmin}_{\omega \in \Omega} \frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} d^2\left(Y_i, \omega\right), \end{aligned} \tag{2}$$

where $L(x, \Pi_\lambda)$ is the leaf node containing x and $N(x, \Pi_\lambda)$ is the number of observations falling in $L(x, \Pi_\lambda)$. If the leaf $L(x, \Pi_\lambda)$ is empty, to avoid ambiguity, we define $0/0 = 0$ by convention and the tree estimate $\hat{m}_{\oplus}(x)$ may take any value in Ω. We call $\hat{m}_{\oplus}(x)$ the tree estimator.

C. Fréchet Mondrian forest

A random forest is an ensemble of multiple random trees. Given $\Pi_{\lambda,M} = \{\Pi_\lambda^{(j)}\}_{j=1}^M$ independently and identically sampled from $\mathrm{MP}(\lambda, [0,1]^p)$, we have the estimator $\hat{m}_{\oplus}^{(j)}(x)$ by (2) based on $\Pi_\lambda^{(j)}$:

$$\begin{aligned} \hat{m}_{\oplus}^{(j)}(x) &= \operatorname*{argmin}_{\omega \in \Omega} \hat{F}_x^{(j)}(\omega) \\ &:= \operatorname*{argmin}_{\omega \in \Omega} \frac{1}{N(x, \Pi_\lambda^{(j)})} \sum_{i: X_i \in L(x, \Pi_\lambda^{(j)})} d^2\left(Y_i, \omega\right). \end{aligned}$$

The forest estimator for $m_{\oplus}(x)$ can be given in two ways. The estimator proposed by [27] adopts the sample Fréchet mean of the tree estimators:

$$\hat{m}_{\oplus,M}(x) = \operatorname*{argmin}_{\omega \in \Omega} \sum_{j=1}^{M} d^2\left(\hat{m}_{\oplus}^{(j)}(x), \omega\right). \tag{3}$$

However, the operation of Fréchet mean does not obey Jensen's inequality [33]. Specifically, for any sequence $\{\omega_1, \ldots, \omega_M\}$ of elements in a general metric space Ω, and for any sequence of real numbers $\{\alpha_1, \ldots, \alpha_M\}$, satisfying $\alpha_j \ge 1$,

$$\begin{aligned} &d^r\left(\operatorname*{argmin}_{\omega \in \Omega} \sum_{j=1}^{M} \alpha_j d^r\left(\omega_j, \omega\right), \omega'\right) \\ \le\, &2^{r-1} \sum_{j=1}^{M} \alpha_j d^r\left(\hat{m}_{\oplus}^{(j)}(x), \omega'\right) \end{aligned}$$

for any positive integer M and $r \ge 1$. Therefore

$$\begin{aligned} \mathbb{E}\left\{d^2\left(\hat{m}_{\oplus,M}(x), m_{\oplus}(x)\right)\right\} \le& 2\sum_{i=1}^{M} \mathbb{E}\left\{d^2\left(\hat{m}_{\oplus}^{(j)}(x), m_{\oplus}(x)\right)\right\} \\ =& 2M \cdot \mathbb{E}\left\{d^2\left(\hat{m}_{\oplus}(x), m_{\oplus}(x)\right)\right\}, \end{aligned}$$

which poses a challenge in demonstrating the theoretical advantage of Fréchet forests over individual Fréchet trees. This also suggests that the theoretical guarantee for the rationality of tree ensembles in non-Euclidean scenarios is not a simple extension of the Euclidean results.

The alternative route introduced in [28] provides a promising direction to advance our research. They use forests to generate local weights rather than directly averaging the results

of the trees. More precisely, the Fréchet forest implicitly constructs a kernel-type weighting function

$$\alpha_{\lambda,M}(x,z)=\frac{1}{M}\sum_{j=1}^{M}\frac{\mathbb{1}_{\{z\in L(x,\Pi_\lambda^{(j)})\}}}{N(x,\Pi_\lambda^{(j)})}, \tag{4}$$

where $L(x,\Pi_\lambda^{(j)})$ is the leaf node containing x of the j-th Mondrian tree based on the partition $\Pi_\lambda^{(j)}$, and $N(x,\Pi_\lambda^{(j)})$ is the number of observations in it. This kernel is equivalent to the empirical probability that x and z share a leaf node within the random forest. With the forest kernel (4), [28] gives the following forest estimator for $m_\oplus(x)$

$$\begin{aligned}\hat{m}_{\oplus,M}(x)&=\operatorname*{argmin}_{\omega\in\Omega}\hat{F}_{x,M}(\omega)\\&:=\operatorname*{argmin}_{\omega\in\Omega}\sum_{i=1}^{n}\alpha_{i,M}(x)d^2\left(Y_i,\omega\right),\end{aligned} \tag{5}$$

where $\alpha_{i,M}(x)=\alpha_{\lambda,M}(x,X_i)$ is used to measure the local contribution of X_i to x. There are two points worth mentioning below. First, let $\alpha_i(x)=\mathbb{1}_{\{X_i\in L(x,\Pi_\lambda)\}}/N(x,\Pi_\lambda)$ be the local weight provided by a Mondrian tree based on the partition Π_λ and $\alpha_i^{(j)}(x)$ be the adaptation of $\alpha_i(x)$ when taking Π_λ as $\Pi_\lambda^{(j)}$, then $\alpha_{i,M}(x)=\frac{1}{M}\sum_{j=1}^{M}\alpha_i^{(j)}(x)$ and

$$\begin{aligned}\hat{F}_{x,M}(\omega)&=\frac{1}{M}\sum_{j=1}^{M}\left(\sum_{i=1}^{n}\alpha_i^{(j)}(x)d^2(Y_i,\omega)\right)\\&=\frac{1}{M}\sum_{j=1}^{M}\hat{F}_x^{(j)}(\omega).\end{aligned} \tag{6}$$

Thus, (5) is equivalent to optimizing the average of the target functions associated with individual Mondrian Fréchet trees. In particular, $\hat{m}_{\oplus,M}(x)$ reduces to the tree estimator $\hat{m}_\oplus(x)$ when $M=1$. Second, When $\Omega=\mathcal{R}$ equipped with the Euclidean distance, $\hat{m}_{\oplus,M}(x)$ has the explicit expression

$$\begin{aligned}\hat{m}_{\oplus,M}(x)&=\sum_{i=1}^{n}\alpha_{i,M}(x)Y_i\\&=\frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N(x,\Pi_\lambda^{(j)})}\sum_{i:X_i\in L(x,\Pi_\lambda^{(j)})}Y_i\right),\end{aligned}$$

which reduces to the classical Euclidean Mondrian forest [21], i.e., the average of all Mondrian tree estimators. This illustrates that (5) serves as a generalization of Euclidean forests for metric space-valued responses. Note that (5) and (3) are two distinct methods, and in the sequel, the forest estimator $\hat{m}_{\oplus,M}(x)$ always refers to (5) based on which the theoretical guarantees are established. In the extreme case $\alpha_{i,M}(x)=0$ for all $1\le i\le n$, namely, no observations fall into the same leaf as x in any Fréchet tree, the situation becomes degenerate and $\hat{m}_{\oplus,M}(x)$ may take any value in Ω.

III. Convergence Analysis

This section develops the theoretical guarantees for the proposed Fréchet Mondrian forests. Our primary objective is to establish the corresponding consistency and non-asymptotic convergence rates under mild geometric and smoothness conditions. These results extend the existing theory for Euclidean responses to general metric space valued responses and demonstrate that random forest constructions remain statistically effective for non-Euclidean regression.

A. Consistency

Consistency is a fundamental requirement for the validity of an estimator. Before presenting the formal results, we first outline the underlying assumptions.

(A1) (Ω,d) is a bounded metric space, i.e., $\operatorname{diam}(\Omega,d)=\sup_{\omega_1,\omega_2\in\Omega}d\left(\omega_1,\omega_2\right)<\infty$.

(A2) X has a bounded positive density $f(x)$ w.r.t. the Lebesgue measure on $[0,1]^p$. In addition, assume that there exists a probability measure μ on Ω satisfying the following property: let $y\to\rho(y|x)$ be the μ-density of Y conditional on $X=x$, for μ-almost all $y\in\Omega$, $\rho(y|x)$ is continuous in x.

(A3) The objects $m_\oplus(x)$ and $\hat{m}_{\oplus,M}(x)$ exist and are unique, the latter almost surely, and, for any $\varepsilon>0$,

$$\inf_{d(\omega,m_\oplus(x))>\varepsilon}\left\{F_x(\omega)-F_x\left(m_\oplus(x)\right)\right\}>0.$$

(A4) $N(x,\Pi_\lambda^{(j)})\to\infty$ for $j=1,\ldots,M$.

Assumptions (A1)–(A3) are commonly used conditions to study the Fréchet regression, see [2]. Assumption (A2) imposes a continuity condition on the conditional distribution of the response. Assumption (A3) is a regularity condition ensuring the consistency of M-estimators (Corollary 3.2.3 of [34]). Assumption (A3) holds automatically in Hadamard spaces (Proposition 4.3 of [35]) when $\mathbb{E}\left\{d^2(Y,\omega)\mid X=x\right\}<\infty$ for some $\omega\in\Omega$, equivalently, when the conditional second moment is finite. Examples of Hadamard spaces include Hilbert spaces, Riemannian manifolds with nonpositive sectional curvature, the Wasserstein space on $\mathcal{R}$, and complete real trees. For other metric spaces, the existence and uniqueness of $m_\oplus(x)$ and $\hat{m}_{\oplus,M}(x)$ can still be guaranteed when the data lie in a geodesically convex subset and the distance function satisfies a suitable convexity condition [8]. For example, on the unit sphere with geodesic distance, the uniqueness of (sample) Fréchet means holds if the support of the underlying distribution is restricted to a hemisphere. For a detailed discussion on Riemannian manifolds, see [36]. Assumption (A4) imposes a condition on the tree growth rate, which is also adopted in [37], [38].

The following result establishes the consistency of Fréchet Mondrian forests: as the sample size grows, the estimators converge to the true regression function. This extends the fundamental consistency property from Euclidean forests to the general metric-space setting.

Theorem 1. *For a fixed $x\in[0,1]^p$ and any $M\ge1$, suppose that (A1)–(A4) hold. If $n\to\infty$ and $\lambda\to\infty$, then $\hat{m}_\oplus(x)$ is pointwise consistent, that is,*

$$d(\hat{m}_{\oplus,M}(x),m_\oplus(x))=o_p(1).$$

B. Convergence rate

Furthermore, we investigate the non-asymptotic convergence rate of the estimators. The error rate for Fréchet regression methods hinges on the complexity of the metric space Ω, which can be characterized by Talagrand's γ_2 measure [39]. This measure is also adopted in [9].

Definition 1. *(i) Given a set $\mathcal{B} \subset \Omega$, an admissible sequence is an increasing sequence $(\mathcal{A}_k)_{k\in\mathcal{N}_0}$ of partitions of $\mathcal{B}$ such that $\mathcal{A}_0 = \{\mathcal{B}\}$ and the cardinality of $\mathcal{A}_k$ is bounded as $\#\mathcal{A}_k \leq 2^{2^k}$ for $k \geq 1$. If every set of $\mathcal{A}_{k+1}$ is contained in a set of $\mathcal{A}_k$, then it is called an increasing sequence of partitions. We denote by $A_k(\omega)$ the unique element of $\mathcal{A}_k$ which contains $\omega \in \mathcal{B}$. (ii) Let $(\mathcal{B}, d)$ be a pseudo-metric space, i.e., d is symmetric, fulfills the triangle inequality, and $d(\omega, \omega) = 0$ for all $\omega \in \mathcal{B}$. Define*

$$\gamma_2(\mathcal{B}, d) := \inf \sup_{\omega\in\mathcal{B}} \sum_{k=0}^{\infty} 2^{\frac{k}{2}} \operatorname{diam}\left(A_k(\omega), d\right),$$

where the infimum is taken over all admissible sequences in $\mathcal{B}$.

Talagrand's γ_2 measure can be bounded by the entropy integral

$$\gamma_2(\mathcal{B}, d) \leq c \int_0^{\infty} \sqrt{\log(N(\mathcal{B}, d, r))}\mathrm{d}r,$$

where $N(\mathcal{B}, d, r)$ is the r-covering number of the set $\mathcal{B}$ w.r.t. the metric d. To establish the risk bound for Fréchet Mondrian trees and forests, we need the following assumptions.

(B1) The object $m_\oplus(x)$ exists. There is $C_{\mathsf{Vlo}} \in [1, \infty)$ such that $C_{\mathsf{Vlo}}^{-1}\, d^2(\omega, m_\oplus(x)) \leq F_x(\omega) - F_x(m_\oplus(x))$ for all $\omega \in \Omega$ and $x \in [0, 1]^p$.

(B2) There are $C_{\mathsf{Ent}} \in [1, \infty)$ and $\alpha \in [1, 2)$ such that, for all $\mathcal{B} \subset \Omega$,

$$\gamma_2(\mathcal{B}, d) \leq C_{\mathsf{Ent}} \max\left(\operatorname{diam}(\mathcal{B}, d), \operatorname{diam}(\mathcal{B}, d)^{\alpha}\right).$$

(B3) There are $\kappa > \frac{2}{2-\alpha}$ and $C_{\mathsf{Mom}} \in [1, \infty)$ such that $\mathbb{E}\left\{d^{\kappa}(Y, m_\oplus(x))\middle| X = x\right\}^{\frac{1}{\kappa}} \leq C_{\mathsf{Mom}}$ for all $x \in [0, 1]^p$.

(B4) Let $C_{\mathsf{Len}} \in [1, \infty)$ such that $\sup_{x_1,x_2\in[0,1]^p} d(m_\oplus(x_1), m_\oplus(x_2)) \leq C_{\mathsf{Len}}$. There exists a probability measure μ on Ω satisfying the following properties: let $C_{\mathsf{Int}} \in [1, \infty)$ such that $\int d^2(y, m_\oplus(x_0))\mu(\mathrm{d}y) \leq C_{\mathsf{Int}}$ for some $x_0 \in [0, 1]^p$; let $y \to \rho(y|x)$ be the μ-density of Y conditional on $X = x$; for μ-almost all $y \in \Omega$, there is $L(y) \geq 0$ and $\beta \in (0, 1]$ such that

$$\left|\rho(y|x) - \rho(y|x')\right| \leq L(y) \cdot \|x - x'\|_2^{\beta}$$

for every $x, x' \in [0, 1]^p$; moreover, there are constants $C_{\mathsf{SmD}}, C_{\mathsf{CdI}} \in [1, \infty)$ such that $\int L(y)^2\mu(\mathrm{d}y) \leq C_{\mathsf{SmD}}^2$ and $\int \rho^2(y|x)\mu(\mathrm{d}y) \leq C_{\mathsf{CdI}}^2$ for all $x \in [0, 1]^p$.

(B5) Define $H(\omega_1, \omega_2) = \left\{\int \left(d(y, \omega_1)+d(y, \omega_2)\right)^2\mu(\mathrm{d}y)\right\}^{\frac{1}{2}}$. There is $C_{\mathsf{Bom}} \in [1, \infty)$ such that, for all $x \in [0, 1]^p$,

$$\mathbb{E}\left\{H(\hat{m}_{\oplus,M}(x), m_\oplus(x))^{\kappa}\middle|\{X_i\}_{i=1}^n, \Pi_{\lambda,M}\right\}^{\frac{1}{\kappa}} \leq C_{\mathsf{Bom}}.$$

The above assumptions were originally introduced by [9] to establish the convergence rate of local Fréchet regression [2] in expectation. Assumption (B1), a quantitative version of Assumption (A3), is a standard condition that regulates the behavior of $F_x(\omega)$ near $m_\oplus(x)$. It is related to the convergence rate of M-estimators and guarantees the uniqueness of the minimizer $m_\oplus(x)$. It always holds when Ω is a Hadamard space (Proposition 4.4 of [35]). Assumption (B2) is satisfied with $\alpha = 1$ for both Euclidean space and any bounded metric space. The bounded Fréchet moment condition in (B3) is a natural extension of the moment condition commonly imposed in Euclidean settings. Assumption (B4) imposes smoothness on $m_\oplus(x)$ via the conditional density $\rho(y|x)$ since the conditional Fréchet mean lacks an explicit expression. Assumption (B5) is a general condition applicable to both bounded metric spaces and Hadamard spaces. For further details, please see Remarks 1 and 2 of [9]. In the next section, we will present four representative examples, including symmetric positive-definite matrices, functional data, compositional data, and probability distributions, all of which satisfy the stated assumptions.

The following theorem establishes a non-asymptotic risk bound for Fréchet Mondrian forests. Remarkably, the convergence rate matches the minimax optimal rate in Euclidean regression under β-Hölder smoothness. This highlights that, despite the lack of linear structure, non-Euclidean forests can achieve comparable efficiency to classical methods.

Theorem 2. *Assume (B1)–(B5) and $M \geq 1$. The mean squared error of the forest estimator $\hat{m}_{\oplus,M}(x)$ with lifetime parameter $\lambda > 0$ satisfies*

$$\begin{aligned}
&\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(x), m_\oplus(x))\right\} \\
&\leq c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{SmD}}\right)^{\frac{2}{2-\alpha}} p^{\frac{\beta}{2-\alpha}} \left(\frac{8}{\lambda^2}\right)^{\beta} + \\
&\quad c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{CdI}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} \mathbb{E}\left(\frac{1}{n\nu_X\left(L(x, \Pi_\lambda)\right)}\right).
\end{aligned}$$

Then the mean integrated squared error is

$$\begin{aligned}
&\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X))\right\} \\
&\leq c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{SmD}}\right)^{\frac{2}{2-\alpha}} p^{\frac{\beta}{2-\alpha}} \left(\frac{8}{\lambda^2}\right)^{\beta} + \\
&\quad c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{CdI}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} \frac{(1+\lambda)^p}{n}.
\end{aligned} \tag{7}$$

In particular, taking $\lambda = \lambda_n \asymp n^{1/(p+2\beta)}$, we have

$$\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X))\right\} \leq O\left(n^{-\frac{2\beta}{p+2\beta}}\right).$$

Theorem 2 holds for Fréchet Mondrian forests with any number of trees. This shows that both Fréchet Mondrian trees ($M = 1$) and forests attain the same convergence rate. The first term on the right-hand side of (7) represents the bias of Fréchet Mondrian forests, caused by the local approximation at the leaves. The second term represents the variance of the Fréchet Mondrian forest, arising from sample estimation. The bound provided here pertains to convergence in expectation and is non-asymptotic, distinguishing it from the results in [2], [28]. In particular, the rate $n^{-2\beta/(p+2\beta)}$ coincides exactly with the minimax rate [22] for nonparametric Euclidean regression under β-Hölder smoothness on $\mathcal{R}^p$. In

contrast to the one-dimensional fixed design studied in [9], Theorem 2 is established under the more general multi-dimensional random design setting. In the random design, the covariates and Mondrian partitions are random, making the local weights stochastic as well. This requires controlling the joint randomness of data and partitions and handling the local effective sample size in each random leaf, including empty-leaf events.

Remark 1. *As in [9], we consider two important metric space classes:*

(1) When Ω is a bounded metric space, Assumption (B2) holds with $\alpha = 1$, Assumption (B3) holds with $C_{Mom} = \operatorname{diam}(\Omega, d)$, and Assumption (B5) holds with $C_{Bom} = 2\operatorname{diam}(\Omega, d)$;

(2) When Ω is a Hadamard space, Assumption (B1) holds with $C_{Vlo} = 1$, and Assumption (B5) holds with $C_{Bom} = c_\kappa C_{Mom} C_{Len} C_{Int}$, provided that $C_{Mom}, C_{Len}, C_{Int}$ all exist.

For a comprehensive treatment on metric geometry, see [40].

IV. Faster convergence of Fréchet forests under higher-Order smoothness

In the previous section, we showed that Fréchet Mondrian forests attain the minimax optimal convergence rate for β-Hölder smooth regression functions in general metric spaces. However, this result alone does not demonstrate any statistical advantage of forests over individual trees, as both estimators share the same convergence rate when $\beta \le 1$. In practical applications, forests often outperform single trees, and this phenomenon also carries over to non-Euclidean settings. Consistent with findings in the Euclidean case [21], we observe that in high-order smooth Fréchet regression, a single Fréchet tree achieves at most the rate corresponding to the Lipschitz case, rendering it suboptimal compared to Fréchet forests. The inherently piecewise constant nature of a single tree limits its ability to fully exploit the higher smoothness of the regression function. In contrast, by averaging the objectives over multiple trees as in (6), forests inherently regularize and smooth the estimator. This aggregation enables forests to adapt more effectively to the underlying smoothness, resulting in a faster convergence rate. To formalize this, we now introduce a strengthened version of the smoothness assumption (B4) on the conditional density $\rho(y|x)$.

(B4′) Further, for every $x, x' \in [0,1]^p$,

$$\|\nabla\rho(y|x) - \nabla\rho(y|x')\|_2 \le L(y) \cdot \|x - x'\|_2^\beta$$

and $\|\nabla\rho(y|x)\|_2 \le L(y)$; X has a positive and C_f-Lipschitz density f w.r.t. the Lebesgue measure on $[0,1]^p$.

The following theorem shows that, under the above high-order smoothness assumption, Fréchet Mondrian forests built with sufficiently many trees continue to exhibit favorable convergence behavior due to bias reduction.

Theorem 3. *Assume (B1)–(B3), (B4′), (B5) and $M \ge 1$. The mean squared error of the forest estimator $\hat{m}_{\oplus,M}(x)$ with lifetime parameter $\lambda > 0$ satisfies*

$$\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(x), m_\oplus(x))\right\}$$

$$\le c_{\alpha,\kappa}\left(C_{Vlo}C_{Bom}C_{Cdl}C_{Mom}C_{Ent}\right)^{\frac{2}{2-\alpha}} \mathbb{E}\left(\frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right) +$$

$$c_{\alpha,\kappa}C_{SmD}^2 C_{Cdl}^{\frac{2\alpha-2}{2-\alpha}}\left(C_{Vlo}C_{Bom}\right)^{\frac{2}{2-\alpha}}\left\{\frac{p}{\lambda^2 M} + \left(\frac{f_1 C_f}{f_0^2}\right)^2 \frac{p^3}{\lambda^4} + \left(\frac{f_1}{f_0}\right)^2 \frac{p^{1+\beta}}{\lambda^{2(1+\beta)}} + \frac{1}{\lambda^2}\sum_{j=1}^{p} e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}\right\},$$

where $f_0 = \inf_{x\in[0,1]^p} f(x)$ and $f_1 = \sup_{x\in[0,1]^p} f(x)$. Set $\varepsilon \in (0, 1/2)$ and $B_\varepsilon = [\varepsilon, 1-\varepsilon]^p$. Then

$$\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X)) \mid X \in B_\varepsilon\right\}$$

$$\le c_{\alpha,\kappa}\left(C_{Vlo}C_{Bom}C_{Cdl}C_{Mom}C_{Ent}\right)^{\frac{2}{2-\alpha}} \frac{1}{f_0(1-2\varepsilon)^p}\frac{(1+\lambda)^p}{n} +$$

$$c_{\alpha,\kappa}C_{SmD}^2 C_{Cdl}^{\frac{2\alpha-2}{2-\alpha}}\left(C_{Vlo}C_{Bom}\right)^{\frac{2}{2-\alpha}}\left\{\frac{p}{\lambda^2 M} + \left(\frac{f_1 C_f}{f_0^2}\right)^2 \frac{p^3}{\lambda^4} + \left(\frac{f_1}{f_0}\right)^2 \frac{p^{1+\beta}}{\lambda^{2(1+\beta)}} + \frac{f_1}{f_0(1-2\varepsilon)^p}\frac{pe^{-\lambda\varepsilon}}{\lambda^3}\right\}.$$

In particular, setting $s = 1+\beta$ and taking $\lambda = \lambda_n \asymp n^{1/(p+2s)}$ and $M = M_n \gtrsim n^{2\beta/(p+2s)}$, we have

$$\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X)) \mid X \in B_\varepsilon\right\} = O\left(n^{-\frac{2s}{p+2s}}\right). \quad (8)$$

In the case where $\varepsilon = 0$, which corresponds to integrating over the whole hypercube, the bound (8) holds if $2s \le 3$. On the other hand, if $2s > 3$, letting $\lambda = \lambda_n \asymp n^{1/(p+3)}$ and $M = M_n \gtrsim n^{1/(p+3)}$ yields

$$\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X))\right\} = O\left(n^{-\frac{3}{p+3}}\right).$$

The theorem again shows that the Fréchet Mondrian forest, with a sufficient number of trees, can achieve the nonparametric optimal rate [22]. However, due to boundary effects, the convergence rate over the entire feature space $[0,1]^p$ deteriorates once $s > 1.5$. This phenomenon is also observed in Euclidean local constant estimators such as Nadaraya–Watson and k-nearest neighbors regressions [41], [42]. Theorem 3 demonstrates that aggregating multiple trees induces locally smoother weights than a single Fréchet tree, thereby better capturing higher-order smooth structures in the Fréchet regression function. In contrast to [9], which improves rates by upgrading Nadaraya–Watson to local linear Fréchet regression, our approach attains similar gains through ensemble aggregation. These represent two distinct routes to exploit higher-order smoothness, with the ensemble effect being notably more subtle in non-Euclidean settings.

From a technical perspective, Theorem 3 differs from the Euclidean analysis of [21], although both rely on the core Mondrian geometry to calibrate bias and variance. In the Euclidean case, a forest prediction is an explicit scalar average of tree outputs, which permits a direct bias–variance decomposition. By contrast, in the non-Euclidean setting, the estimator is defined implicitly as a minimizer of a squared-distance functional, so the entire risk analysis must be rebuilt within the M-estimation framework in combination with empirical-process techniques. Moreover, as classical Jensen's inequality

fails in general metric spaces, we average objectives (squared-distance functionals) via forest-induced weights, shifting ensembling from averaging point estimates to averaging losses. All results are developed under general metric spaces, without relying on any linear structure. We instantiate the theory on bounded metric spaces and Hadamard spaces. Consequently, [21] emerges as a special case of our framework, and the provable ensemble advantage extends beyond Euclidean domains. Below, we illustrate through several examples that a single tree cannot attain the convergence rate provided by Theorem 3.

a) Symmetric positive-definite matrices: Symmetric positive-definite matrices frequently arise in optimization, machine learning, and geometry, particularly in the context of Riemannian metrics and covariance structures. To define commonly used metrics between such matrices, some preliminary concepts are needed. For a matrix S, let $\lfloor S \rfloor$ denote its strictly lower triangular matrix, and $\mathbb{D}(S)$ denote its diagonal part. Let I_m denote the $m \times m$ identity matrix. For a symmetric matrix $A \in \mathcal{R}^{m\times m}$, the matrix exponential is defined by $\exp(A) = I_m + \sum_{j=1}^{\infty} \frac{1}{j!} A^j$. Conversely, for a symmetric positive-definite matrix S, the matrix logarithm is defined as $\log(S) = A$ such that $\exp(A) = S$. If S is symmetric positive-definite, there exists a unique lower triangular matrix P with positive diagonal entries such that $PP^\top = S$. This matrix P is called the Cholesky factor of S, denoted by $\mathscr{L}(S)$. Given two symmetric positive-definite matrices S_1 and S_2, the log-Euclidean metric [43] between them is defined as

$$d(S_1, S_2) = \| \log(S_1) - \log(S_2) \|_{\mathrm{F}}.$$

The log-Cholesky metric [44] is defined as

$$d(S_1, S_2) = d_{\mathcal{L}}(\mathscr{L}(S_1), \mathscr{L}(S_2)), \tag{9}$$

where $d_{\mathcal{L}}(P_1, P_2) = \{\|\lfloor P_1 \rfloor - \lfloor P_2 \rfloor\|_{\mathrm{F}}^2 + \|\log \mathbb{D}(P_1) - \log \mathbb{D}(P_2)\|_{\mathrm{F}}^2\}^{1/2}$. And the affine-invariant metric [45], [46] is defined as

$$d(S_1, S_2) = \left\| \log\left(S_1^{-1/2} S_2 S_1^{-1/2} \right) \right\|_{\mathrm{F}}.$$

The space of symmetric positive-definite matrices, equipped with the Frobenius, log-Euclidean, log-Cholesky, or affine-invariant metric, forms a Riemannian manifold of nonpositive sectional curvature. We begin with an example of Fréchet regression, where the responses are symmetric positive-definite matrices endowed with the log-Euclidean metric, to show that a single Fréchet Mondrian tree does not attain the convergence rate achieved by forests.

Example 1. *Let $\mathcal{S}_2^+$ be the collection of 2×2 symmetric positive-definite matrices. We equip $\mathcal{S}_2^+$ with the log-Euclidean metric. Let (X, Y) be a $[0,1] \times \mathcal{S}_2^+$-valued random pair with the following relationship*

$$Y = \begin{pmatrix} \frac{e^{1+f(X,\varepsilon)} + e^{1-f(X,\varepsilon)}}{2} & \frac{e^{1+f(X,\varepsilon)} - e^{1-f(X,\varepsilon)}}{2} \\ \frac{e^{1+f(X,\varepsilon)} - e^{1-f(X,\varepsilon)}}{2} & \frac{e^{1+f(X,\varepsilon)} + e^{1-f(X,\varepsilon)}}{2} \end{pmatrix},$$

where $f(X, \varepsilon) = X + 1 + \varepsilon$ with $X \sim U([0,1])$ and the random noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of X. A single Fréchet Mondrian tree $\hat{m}_{\oplus}(x)$ is used for Fréchet regression. Here we stipulate that the tree returns $e \cdot I_2$ if the leaf $L(x, \Pi_\lambda)$ is empty.

b) Functional data: Functional data [47], [48] consists of curves, surfaces, or other continuous objects that vary over a domain, typically representing measurements collected over time or space. In functional data analysis, most works focus on $L^2([0,1])$, that is, $L^2(E, \mathscr{B}, \xi)$ space with $E = [0,1]$, $\mathscr{B}$ the Borel σ-field of $[0,1]$ and ξ Lebesgue measure. Specifically, $L^2([0,1])$ is the collection of measurable functions f on $[0,1]$ that satisfy $\int_0^1 |f(t)|^2 \mathrm{d}t < \infty$. The vector space operations of addition and scalar multiplication are defined by $(f \oplus g)(t) = f(t) + g(t)$ and $(c \odot f)(t) = c \cdot f(t)$ for any $c \in \mathcal{R}$ and $f, g \in L^2([0,1])$. Further, $f \ominus g = f \oplus (-1 \odot g)$. It is well known that $L^2([0,1])$ is a separable Hilbert space under the inner product $\langle f, g \rangle = \int_0^1 f(t) g(t) \mathrm{d}t$. Below is an example of Fréchet regression with functional responses, where a single tree cannot achieve the convergence rate of forests.

Example 2. *Let $L^2([0,1])$ be the collection of square integrable functions on $[0,1]$. We equip $L^2([0,1])$ with $d(f, g) = \langle f \ominus g, f \ominus g \rangle^{1/2}$ induced by the inner product. Let (X, Y) be a $[0,1] \times L^2([0,1])$-valued random pair with the following relationship*

$$Y = f_0 \oplus \{(X + \varepsilon) \odot g_0\},$$

where $f_0, g_0 \in L^2([0,1])$ with $g_0 \neq 0$, $X \sim U([0,1])$, and the random noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent of X. A single Fréchet Mondrian tree $\hat{m}_{\oplus}(x)$ is used for Fréchet regression. Here we stipulate that the tree returns $f_0 \ominus g_0$ if the leaf $L(x, \Pi_\lambda)$ is empty.

c) Compositional data: Compositional data [49], [50] consists of vectors whose components represent proportions of a whole and convey relative, rather than absolute, information. Such data naturally arise in various fields, including geochemistry, economics, and microbiome studies, and require specialized statistical methods to account for their inherent constraints. Mathematically, compositional data lie in the open $(k-1)$-dimensional simplex:

$$\mathcal{U}^{k-1} = \left\{ u = (u_{(1)}, u_{(2)}, \ldots, u_{(k)})^\top \in \mathcal{R}_+^k : \sum_{i=1}^k u_{(i)} = 1 \right\}.$$

The simplex is endowed with an Aitchison geometry, which turns it into a separable Hilbert space with operations defined in the log-ratio framework. For any $u, v \in \mathcal{U}^{k-1}$, the additive operation and scalar multiplication are given by

$$u \oplus v = \left(\frac{u_{(1)} v_{(1)}}{\sum_{i=1}^k u_{(i)} v_{(i)}}, \frac{u_{(2)} v_{(2)}}{\sum_{i=1}^k u_{(i)} v_{(i)}}, \ldots, \frac{u_{(k)} v_{(k)}}{\sum_{i=1}^k u_{(i)} v_{(i)}} \right)^\top$$

$$c \odot u = \left(\frac{u_{(1)}^c}{\sum_{i=1}^k u_{(i)}^c}, \frac{u_{(2)}^c}{\sum_{i=1}^k u_{(i)}^c}, \ldots, \frac{u_{(k)}^c}{\sum_{i=1}^k u_{(i)}^c} \right)^\top.$$

The associated inner product in the Aitchison geometry is

$$\langle u, v \rangle = \frac{1}{2k} \sum_{i=1}^k \sum_{j=1}^k \log \frac{u_{(i)}}{u_{(j)}} \log \frac{v_{(i)}}{v_{(j)}}.$$

The induced Aitchison metric is given by

$$d(u, v) = \sqrt{\sum_{i=1}^k \left(\log \frac{u_{(i)}}{h(u)} - \log \frac{v_{(i)}}{h(v)} \right)^2},$$

where $h(u) = \left(\prod_{i=1}^{k} u_{(i)}\right)^{1/k}$ is the geometric mean of u. We next present an example with compositional data as responses, illustrating the advantages of forests over trees.

Example 3. *Let $\mathcal{U}^2$ be the collection of three-dimensional compositional data. We equip $\mathcal{U}^2$ with the Aitchison metric. Let (X, Y) be a $[0,1]\times\mathcal{U}^2$-valued random pair with the following relationship*

$$Y = \left(\frac{1}{f(X,\varepsilon_1,\varepsilon_2)}, \frac{e^{X+1+\varepsilon_1}}{f(X,\varepsilon_1,\varepsilon_2)}, \frac{e^{2X+2+\varepsilon_2}}{f(X,\varepsilon_1,\varepsilon_2)}\right)^\top$$

with $f(X,\varepsilon_1,\varepsilon_2) = 1 + e^{X+1+\varepsilon_1} + e^{2X+2+\varepsilon_2}$, $X \sim U([0,1])$, and the random noise $\varepsilon_1, \varepsilon_2 \sim \mathcal{N}(0,\sigma^2)$ are independent of X. A single Fréchet Mondrian tree $\hat{m}_\oplus(x)$ is used for Fréchet regression. Here we stipulate that the tree returns $(1/3, 1/3, 1/3)^\top$ if the leaf $L(x,\Pi_\lambda)$ is empty.

d) Probability distributions: The Wasserstein space [51], [52] is a metric space of probability measures equipped with the Wasserstein distance, which arises naturally in optimal transport theory. It has become an important tool in statistics and machine learning, with applications ranging from generative modeling and domain adaptation to distributional data analysis and shape matching. Formally, let $(\mathcal{M}, d_\mathcal{M})$ be a Polish metric space. The k-Wasserstein space on $\mathcal{M}$ is

$$\mathcal{W}_k(\mathcal{M}) = \left\{\xi \in \mathcal{P}(\mathcal{M}) : \int d^k_\mathcal{M}(z, z_0)\,\xi(\mathrm{d}z) < \infty, z_0 \in \mathcal{M}\right\},$$

where $\mathcal{P}(\mathcal{M})$ is the set of Borel probability measures on $\mathcal{M}$. The Wasserstein distance between two measures $\xi_1, \xi_2 \in \mathcal{W}_k(\mathcal{M})$ is defined as

$$d(\xi_1,\xi_2) = \left\{\inf_{\gamma\in\Gamma(\xi_1,\xi_2)} \int_{\mathcal{M}\times\mathcal{M}} d^k_\mathcal{M}(z_1,z_2)\gamma(\mathrm{d}(z_1,z_2))\right\}^{1/k},$$

where $\Gamma(\xi_1,\xi_2)$ denotes the set of all couplings (transport plans) between ξ_1 and ξ_2. The space $\mathcal{W}_k(\mathcal{M})$ is a complete geodesic metric space; however, unlike the three spaces discussed above, it is not always a Hadamard space (Section 4 of [53]). In what follows, we consider the most commonly used Wasserstein space, $\mathcal{W}_2(\mathcal{R})$, comprising all probability measures on $\mathcal{R}$ with finite second moments. This space is a Hadamard space, and in this case, the Wasserstein distance has an analytic solution:

$$d(\xi_1,\xi_2) = \left\{\int_0^1 (Q_1(t) - Q_2(t))^2\,\mathrm{d}t\right\}^{1/2}, \qquad (10)$$

where Q_1 and Q_2 are the quantile functions corresponding to ξ_1 and ξ_2, respectively. Below is another example of Fréchet regression with distributional responses.

Example 4. *Let $\mathcal{W}_2(\mathcal{R})$ be the collection of probability distributions on $\mathcal{R}$ with finite second moments. We equip $\mathcal{W}_2(\mathcal{R})$ with the Wasserstein distance. Let (X, Y) be a $[0,1]\times\mathcal{W}_2(\mathcal{R})$-valued random pair with the following relationship*

$$Y = \mathcal{N}(X + 1 + \varepsilon, 1),$$

where $X \sim U([0,1])$ and the random noise $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ is independent of X. A single Fréchet Mondrian tree $\hat{m}_\oplus(x)$ is used for Fréchet regression. Here we stipulate that the tree returns $\mathcal{N}(0,1)$ if the leaf $L(x,\Pi_\lambda)$ is empty.

The following proposition provides a lower bound for the convergence rate of Fréchet Mondrian trees in the preceding examples.

Proposition 1. *In Example 1–4, Assumptions (B1)–(B3), (B4′) with $\beta = 1$, and (B5) hold, and there exist absolute constants C_1, C_2 such that the Fréchet Mondrian tree estimator satisfies, for any $n \geq 18$,*

$$\inf_{\lambda\in\mathcal{R}_+} \mathbb{E}\left\{d^2(\hat{m}_\oplus(X), m_\oplus(X))\right\} \geq C_1 \wedge C_2 \left(\frac{\sigma^2}{n}\right)^{2/3},$$

where the constants vary in different examples.

According to Theorem 3, a Fréchet Mondrian forest with sufficiently many trees can attain the convergence rate of $n^{-3/4}$ with $s = 2$ and $p = 1$. However, the proposition above shows that, for Example 1–4, a single Fréchet Mondrian tree can achieve a rate of $n^{-2/3}$ at most. This underscores the advantages of Fréchet forests over single Fréchet trees and highlights the statistical benefits of ensemble aggregation.

V. Numerical Study

In this section, we investigate three Fréchet regression settings where the response variable takes values in probability distributions, symmetric positive-definite matrices and spherical data. We compare the performance of Fréchet Mondrian trees (FMT) and forests (FMF), while including global Fréchet regression (GFR) [2] and local linear Fréchet regression (LFR) [2] as benchmarks. GFR and LFR can be implemented via the `frechet` R package [54]. Across all simulation settings, performance is evaluated by the average mean squared error (AMSE) over 100 Monte Carlo replicates. For example, for the rth replicate, $\hat{m}^r_{\oplus,M}(x)$ denotes the prediction given by the FMF estimator $\hat{m}_{\oplus,M}(x)$, where $M = 1$ corresponds to FMT; the prediction accuracy of FMF is quantitatively assessed using the mean squared error (MSE):

$$\mathrm{MSE}_r(\hat{m}_{\oplus,M}) = \frac{1}{1000}\sum_{i=1}^{1000} d^2(\hat{m}^r_{\oplus,M}(X_i), m_\oplus(X_i)),$$

computed over an independent testing set $\{X_i\}_{i=1}^{1000}$. The final performance evaluation for FMF uses the AMSE, obtained by averaging the MSE over 100 simulation runs.

According to the theoretical analysis, the tree depth (controlled by the parameter $\lambda \in \mathcal{R}_+$) increases with the training sample size. However, as the exact relationship is unknown in practice, tuning λ remains challenging. To this end, we reformulate the stopping criterion in terms of the maximum number of samples allowed in a leaf node, denoted as k. Specifically, a node is not split further if the number of samples within it does not exceed k. When the training sample size is 100, the default value of k is set to 3. Thus, for a general sample size n, we set $k = \lfloor 2(n/100)^{3/(p+3)}\rfloor + 1$. The exponent $3/(p+3)$ is derived by balancing the relation $n \asymp (1+\lambda)^p \cdot k$, where $(1+\lambda)^p$ is the expected number of leaves in a Mondrian tree (see Proposition 2 of [21]) and

$\lambda \asymp n^{1/(p+3)}$ corresponds to the optimal tuning in Theorem 3 under smoothness $s = 2$. This empirical rule yields high-performance forests in the simulations below.

A. *Fréchet regression for probability distributions*

First, we consider a Fréchet regression setting with distributional responses, serving as a representative case of Examples 1–4. Let $\mathcal{W}_2(\mathcal{R})$, equipped with the Wasserstein distance (10), denote the space of probability distributions on $\mathcal{R}$ with finite second moments. Independent covariates $X_1, \ldots, X_n$ are drawn from $U([0,1]^p)$, and the response $Y \in \mathcal{W}_2(\mathcal{R})$ is generated as

$$Y = \mathcal{N}(\mu_Y, \sigma_Y^2),$$

where the mean parameter μ_Y is drawn from

$$\mu_Y \sim \mathcal{N}\big(\sin(4\pi\beta_1^\top X)\,(2\beta_2^\top X - 1),\, 0.2^2\big)$$

and the variance parameter is given by $\sigma_Y = 2\left(X_{(1)} - X_{(2)}\right)$. We consider two situations: (i) $p = 2$: $\beta_1 = (0.75, 0.25)^\top$, $\beta_2 = (0.25, 0.75)^\top$; (ii) $p = 5$: $\beta_1 = (0.1, 0.2, 0.3, 0.4, 0)^\top$, $\beta_2 = (0, 0.1, 0.2, 0.3, 0.4)^\top$. The above setting satisfies Assumptions (B1)–(B3), (B4′) with $\beta = 1$, and (B5).

We consider training sample sizes $\{10^2, 10^{2.4}, 10^{2.8}, \ldots, 10^4\}$ (rounded down to the nearest integer when necessary) and vary the forest size (i.e., number of trees) over $\{1, 3, 5, \ldots, 35\}$. For each configuration, we compute the base-10 logarithm of the corresponding AMSE; the results are displayed in Fig. 1. The observed patterns are similar for both input dimensions $p = 2$ and $p = 5$, and we therefore focus on the case $p = 2$ (left panel) for illustration. First, for a fixed training sample size, the AMSE decreases as the number of trees increases. This indicates that, under high-order smoothness, Fréchet Mondrian forests benefit from larger ensemble sizes, exhibiting a clear advantage over a single tree. However, the decreasing trend of testing errors gradually levels off, suggesting that beyond a certain threshold, additional trees yield only marginal improvements. This phenomenon aligns with the theoretical result that the forest's risk has effectively reached its optimal convergence rate. Second, as the training sample size increases, both Fréchet Mondrian trees and forests exhibit decreasing testing errors. Moreover, the flattening point of the error curve shifts to the right, indicating that a larger number of trees is required for the forest to attain optimal performance. This again supports the theoretical findings, which suggest that the threshold number of trees needed for optimal convergence increases with the training sample size.

To further evaluate the empirical performance of our approach, we compare FMT and FMF with 15, 35, and 500 trees (FMF15, FMF35, and FMF500, respectively) against GFR and LFR. The training sample size varies over $\{100, 500, 1000\}$. Owing to the high computational cost of LFR, all methods are evaluated on an independent testing set of size 100. The results are reported in Table I. As the true Fréchet regression function is nonlinear, GFR, as an extension of Euclidean linear regression, performs the worst. Moreover, its performance does not improve substantially with larger training samples. In contrast, LFR performs better than GFR owing to its local linear kernel modeling. However, for higher-dimensional predictors ($p = 5$), LFR is unavailable due to implementation constraints in the `frechet` package, which supports predictor dimensions below three. When $p = 2$, LFR still performs worse than FMF and even underperforms FMT as the sample size increases. The performance of FMT improves markedly through the ensemble of multiple trees, yet the benefit tapers off when the number of trees exceeds 35. As the number of training samples increases, the mean and standard deviation of the MSE decrease for all methods.

We also investigate the influence of the leaf-size parameter k, which determines the maximum sample count permitted in each leaf node. In the simulations, k is set by default to $\lfloor 2(n/100)^{3/(p+3)} \rfloor + 1$. To assess the sensitivity of model performance to this parameter, we train both FMT and FMF of 35 trees while varying k around its default value. Taking $p = 2$ as an illustrative example, Fig. 2 depicts the testing errors as functions of k for $n = 500$ and $n = 1000$. The resulting

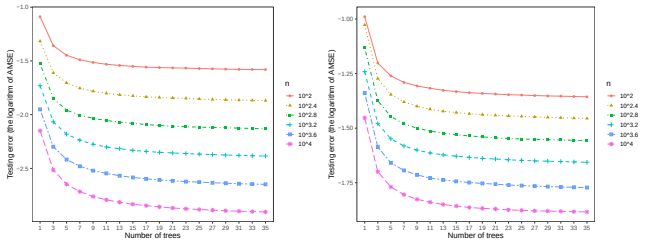

Fig. 1. Testing errors of Fréchet Mondrian forests with varying tree counts for $p = 2$ (left) and $p = 5$ (right) with distributional responses.

TABLE I
AVERAGE MSE (STANDARD DEVIATION) OF DIFFERENT METHODS FOR DISTRIBUTIONAL RESPONSES OVER 100 SIMULATION RUNS. BOLD-FACED NUMBERS INDICATE THE BEST PERFORMERS.

	n	GFR	LFR	FMT	FMF35	FMF500
$p = 2$	100	0.310	0.074	0.113	0.034	**0.032**
		(0.029)	(0.025)	(0.038)	(0.009)	(0.008)
	500	0.300	0.055	0.042	0.010	**0.009**
		(0.028)	(0.023)	(0.011)	(0.002)	(0.002)
	1000	0.298	0.051	0.031	0.006	**0.006**
		(0.027)	(0.023)	(0.008)	(0.002)	(0.001)
$p = 5$	100	0.249	NA	0.295	0.136	**0.132**
		(0.026)		(0.040)	(0.014)	(0.014)
	500	0.237	NA	0.216	0.082	**0.078**
		(0.022)		(0.034)	(0.009)	(0.009)
	1000	0.235	NA	0.170	0.066	**0.063**
		(0.022)		(0.024)	(0.008)	(0.007)

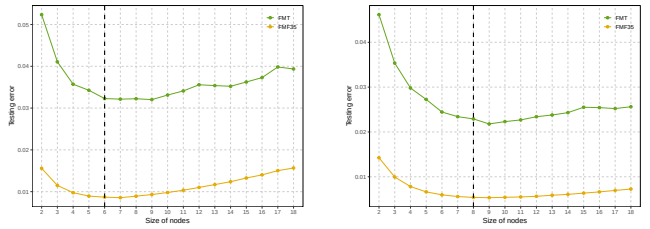

Fig. 2. Testing errors of Fréchet Mondrian trees and forests with varying nodesize k for $n = 500$ (left) and $n = 1000$ (right) with distributional responses. The black vertical dashed line marks the default choice for k.

U-shaped error curves reflect the typical bias–variance trade-off: when k is too small, the estimators exhibit high variance; conversely, when k is too large, bias becomes dominant. The black dashed line in each panel indicates the default choice of k, which lies close to the minimum of the U-shaped curve. This observation suggests that the default choice provides a reasonable balance between bias and variance.

B. Fréchet regression for symmetric positive-definite matrices

As another representative case, we study Fréchet regression with responses being symmetric positive-definite matrices. An $m \times m$ symmetric matrix A is said to follow a matrix-variate normal distribution, denoted by $\mathcal{N}_{mm}(M; \sigma^2)$ [26], if $A = \sigma Z + M$, where M is an $m \times m$ symmetric matrix. The random matrix Z has independent diagonal entries distributed as $\mathcal{N}(0,1)$ and independent off-diagonal entries distributed as $\mathcal{N}(0, 1/2)$. We consider the following setting, which satisfies Assumptions (B1)–(B3), (B4′) with $\beta = 1$ and (B5).

Let $\mathcal{S}_3^+$ be the space of 3×3 symmetric positive-definite matrices endowed with the log-Cholesky metric (9). We independently generate $X_1, \ldots, X_n$ from the uniform distribution $U([0,1]^p)$. The response $Y \in \mathcal{S}_3^+$ is generated as

$$\log(Y) \sim \mathcal{N}_{33}\left(f(X); 0.2^2\right)$$

with

$$f(X) = \begin{pmatrix} 1 & \rho_1(X) & \rho_2(X) \\ \rho_1(X) & 1 & \rho_1(X) \\ \rho_2(X) & \rho_1(X) & 1 \end{pmatrix}$$

and $\rho_1(X) = 0.8\cos\left(4\pi\beta_1^\top X\right), \rho_2(X) = 0.4\cos\left(4\pi\beta_2^\top X\right)$. The choice of (β_1, β_2) for $p = 2$ and $p = 5$ follows that in Section V-A.

Fig. 3 summarizes the effects of sample size and the number of trees on the performance of FMF, following the same setup as before. The observed trends are consistent with those in the previous experiment and are therefore not repeated here. We next compare the competing methods for symmetric positive-definite matrix responses, keeping all other settings unchanged. The results, summarized in Table II, are similar to those observed for distributional responses, except that for $p = 2$, LFR achieves the best performance, while FMF performs competitively. The superiority of forests over individual trees remains evident, which confirms that the ensemble benefits of FMF persist across different metric spaces. Finally, taking $p = 2$ as an example, we again examine the sensitivity of FMF to the leaf-size parameter k. The results in Fig. 4 exhibit a pattern similar to that observed in the previous experiment.

C. Fréchet regression for spherical data

Finally, we examine Fréchet regression with spherical responses. Let $\mathbb{S}^2$ denote the unit sphere in $\mathcal{R}^3$, endowed with the geodesic distance. For two points $s_1, s_2 \in \mathbb{S}^2$, the geodesic distance between them is defined by

$$d(s_1, s_2) = \arccos\left(s_1^\top s_2\right).$$

The unit sphere $\mathbb{S}^2$ is a bounded Riemannian manifold with positive curvature and therefore satisfies Assumptions (B2),

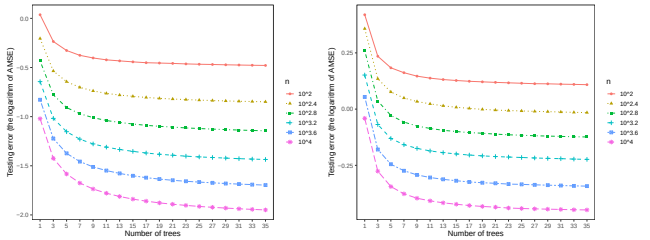

Fig. 3. Testing errors of Fréchet Mondrian forests with varying tree counts for $p = 2$ (left) and $p = 5$ (right) with matrix responses.

TABLE II
AVERAGE MSE (STANDARD DEVIATION) OF DIFFERENT METHODS FOR MATRIX RESPONSES OVER 100 SIMULATION RUNS. BOLD-FACED NUMBERS INDICATE THE BEST PERFORMERS.

	n	GFR	LFR	FMT	FMF35	FMF500
$p = 2$	100	1.925	**0.253**	1.038	0.321	0.298
		(0.154)	(0.207)	(0.209)	(0.055)	(0.049)
	500	1.885	**0.054**	0.404	0.083	0.074
		(0.148)	(0.009)	(0.072)	(0.010)	(0.010)
	1000	1.876	**0.032**	0.281	0.050	0.043
		(0.147)	(0.007)	(0.062)	(0.006)	(0.005)
$p = 5$	100	2.083	NA	2.664	1.326	**1.278**
		(0.137)		(0.436)	(0.122)	(0.110)
	500	1.975	NA	1.927	0.837	**0.808**
		(0.149)		(0.287)	(0.081)	(0.074)
	1000	1.963	NA	1.581	0.688	**0.666**
		(0.143)		(0.231)	(0.072)	(0.060)

(B3), and (B5), as noted in Remark 1. Assumption (B1), however, is generally violated on $\mathbb{S}^2$, though it may hold under specially designed distributions, such as the contracted uniform distribution on the sphere discussed in [9]. We now consider the following setting, regardless of whether Assumption (B1) holds, to observe how the performance of Fréchet Mondrian forests evolves as the number of trees increases with spherical responses.

We independently generate covariates $X_1, \ldots, X_n \sim U([0,1]^p)$. Let the Fréchet regression function be

$$m_\oplus(X) = \begin{pmatrix} \{1 - (\beta_1^\top X)^2\}^{1/2} \cos(\pi \beta_2^\top X) \\ \{1 - (\beta_1^\top X)^2\}^{1/2} \sin(\pi \beta_2^\top X) \\ \beta_1^\top X \end{pmatrix},$$

which maps each X onto the unit sphere $\mathbb{S}^2$. We generate bivariate normal noise ε_i on the tangent space $T_{m_\oplus(X_i)}\mathbb{S}^2$, and then map ε_i back to $\mathbb{S}^2$ using the Riemannian exponential

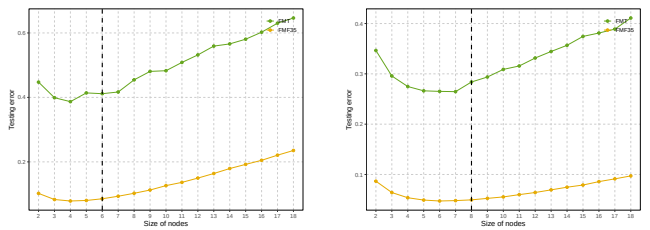

Fig. 4. Testing errors of Fréchet Mondrian trees and forests with varying nodesize k for $n = 500$ (left) and $n = 1000$ (right) with matrix responses. The black vertical dashed line marks the default choice for k.

map to obtain the response Y_i. Specifically, we independently draw $\delta_{i1}, \delta_{i2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.2^2)$ and set $\varepsilon_i = \delta_{i1} v_1 + \delta_{i2} v_2$, where $\{v_1, v_2\}$ is an orthonormal basis of the tangent space $T_{m_\oplus(X_i)}\mathbb{S}^2$. The response Y_i is then generated via

$$\begin{aligned} Y_i &= \operatorname{Exp}_{m_\oplus(X_i)}(\varepsilon_i) \\ &= \cos(\|\varepsilon_i\|_2)\, m_\oplus(X_i) + \sin(\|\varepsilon_i\|_2)\, \frac{\varepsilon_i}{\|\varepsilon_i\|_2}, \end{aligned}$$

where $\operatorname{Exp}_{m_\oplus(X_i)}$ denotes the Riemannian exponential map on $\mathbb{S}^2$ at $m_\oplus(X_i)$. We consider two situations: (i) $p = 2$: $\beta_1 = (1, 0)^\top$, $\beta_2 = (0, 1)^\top$; (ii) $p = 5$: $\beta_1 = (0.1, 0.2, 0.3, 0.4, 0)^\top$, $\beta_2 = (0, 0.1, 0.2, 0.3, 0.4)^\top$.

We adopt the same configurations for training sample sizes and numbers of trees as in the previous experiments, and the corresponding results are presented in Fig. 5. As the results are analogous to those for probability distributions and symmetric positive-definite matrices, repeated descriptions are omitted. All consistent pattern highlights the superiority of Fréchet forests over individual trees and confirms that ensemble aggregation markedly improves the accuracy of Fréchet regression.

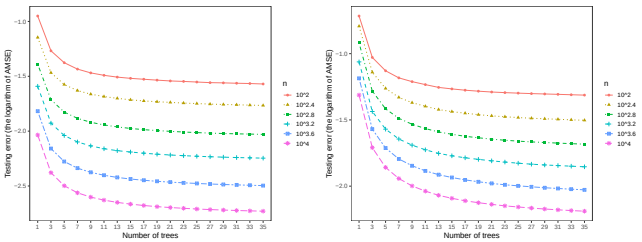

Fig. 5. Testing errors of Fréchet Mondrian forests with varying tree counts for $p = 2$ (left) and $p = 5$ (right) with spherical data responses.

VI. Discussion

Under the Fréchet regression framework with responses taking values in a general metric space, this paper provides the first non-asymptotic analysis of Fréchet random forests by establishing a finite-sample prediction risk bound. We show that, with appropriate tuning, Fréchet Mondrian forests attain the same convergence rate as their Euclidean counterparts. Moreover, we identify conditions under which Fréchet random forests outperform individual trees through bias reduction, thereby highlighting the ensemble advantage in non-Euclidean settings. Collectively, these results contribute to a deeper theoretical understanding of Fréchet random forests. An important insight is that averaging the objective function, rather than the estimators themselves, can also lead to improved statistical efficiency.

The favorable geometric properties of the Mondrian process [21] have been crucial in developing the theoretical guarantees presented herein. Recent advances on random tessellation processes and Tessellation forests [24] suggest promising directions for further generalization. By leveraging the stochastic geometry underlying such processes, one can extend the present framework to define Fréchet Tessellation trees and forests and establish their convergence properties under varying smoothness assumptions. This would move from axis-aligned to oblique partitioning, providing a broader foundation for non-Euclidean random forests and enriching their theoretical and methodological development.

Acknowledgments

The authors are very grateful to the Editor, Associate Editor, and anonymous reviewers for their valuable comments that improved the quality of this article.

Appendix A
Proof of Theorem 1

Proof. For a fix $x \in [0,1]^p$, by Corollary 3.2.3 of [34] and Assumption (A3), we only need to prove $\sup_{\omega\in\Omega}\left|\hat{F}_{x,M}(\omega) - F_x(\omega)\right|$ to zero in probability. To do this, we show $\hat{F}_{x,M}(\cdot) \rightsquigarrow F_x(\cdot)$ in $l^\infty(\Omega)$ and apply Theorem 1.3.6 of [34]. Thanks to Theorem 1.5.4 of [34], this weak convergence is equivalent to $\hat{F}_{x,M}(\cdot)$ being asymptotically tight and the marginals converging weakly. Furthermore, by Theorem 1.5.7 of [34], this asymptotic tightness holds if $\hat{F}_{x,M}(\omega)$ is asymptotically tight in $\mathcal{R}$ for every $\omega \in \Omega$, and $\hat{F}_{x,M}(\cdot)$ is asymptotically uniformly d-equicontinuous in probability. Thus, the proof will be finished if the following conditions hold.

(i) $\hat{F}_{x,M}(\omega) - F_x(\omega) = o_p(1)$ for each $\omega \in \Omega$,

(ii) For all $\varepsilon, \eta > 0$, there exists $\delta > 0$ such that

$$\limsup_n \mathbb{P}\left\{\sup_{d(\omega_1,\omega_2)<\delta}\left|\hat{F}_{x,M}(\omega_1) - \hat{F}_{x,M}(\omega_2)\right| > \varepsilon\right\} < \eta.$$

First, prove (i): Let

$$\tilde{F}_{x,M}(\omega) = \frac{1}{M}\sum_{j=1}^{M}\mathbb{E}\Big\{d^2(Y,\omega) \mid X \in L(x, \Pi_\lambda^{(j)})\Big\}.$$

Then we consider decomposing $\hat{F}_{x,M}(\omega) - F_x(\omega)$ into a variance-type term $\hat{F}_{x,M}(\omega) - \tilde{F}_{x,M}(\omega)$ and a bias-type term $\tilde{F}_{x,M}(\omega) - F_x(\omega)$. We will show that both terms converge to zero in probability for any $\omega \in \Omega$.

We first prove $\hat{F}_{x,M}(\omega) - \tilde{F}_{x,M}(\omega) \xrightarrow{p} 0$. For convenience in notation, let $\alpha_i^{(j)}(x) = \mathbb{1}_{\{X_i \in L(x, \Pi_\lambda^{(j)})\}}$, indicating that a training point X_i belongs to the same leaf node as x in the j-th Mondrian tree based on the partition $\Pi_\lambda^{(j)}$. By Assumption (A4), the number of training points falling in a leaf node goes to infinity. If we condition on the variables $\alpha_i^{(j)}(x)$, the subset of the training data falling in $L(x, \Pi_\lambda^{(j)})$ is independent and identically distributed in the rectangle $L(x, \Pi_\lambda^{(j)})$. For $1 \le j \le M$, the weak law of large numbers gives

$$\begin{aligned} &\frac{1}{N(x, \Pi_\lambda^{(j)})}\sum_{i=1}^{n}\alpha_i^{(j)}(x) d^2(Y_i, \omega) \\ &- \mathbb{E}\Big\{d^2(Y,\omega) \mid X \in L(x, \Pi_\lambda^{(j)})\Big\} \xrightarrow{p} 0. \end{aligned}$$

Therefore, averaging over total M Mondrian trees, we get

$$\begin{aligned}
&\hat{F}_{x,M}(\omega) - \tilde{F}_{x,M}(\omega) \\
&= \sum_{i=1}^{n} \frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{X_i \in L(x,\Pi_\lambda^{(j)})\}}}{N(x,\Pi_\lambda^{(j)})} d^2(Y_i,\omega) \\
&\quad - \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}\left\{ d^2(Y,\omega) \mid X \in L(x,\Pi_\lambda^{(j)}) \right\} \\
&= \frac{1}{M} \sum_{j=1}^{M} \left[\frac{1}{N(x,\Pi_\lambda^{(j)})} \sum_{i=1}^{n} \alpha_i^{(j)}(x) d^2(Y_i,\omega) \right. \\
&\quad \left. - \mathbb{E}\left\{ d^2(Y,\omega) \mid X \in L(x,\Pi_\lambda^{(j)}) \right\} \right] \\
&\xrightarrow{p} 0.
\end{aligned} \tag{11}$$

Now we turn to prove $\tilde{F}_{x,M}(\omega) - F_x(\omega) \xrightarrow{p} 0$. For any $\omega \in \Omega$, then

$$\mathbb{E}\{ d^2(Y,\omega) \mid X = x \} = \int_\Omega d^2(y,\omega) \rho(y|x) \mu(\mathrm{d}y).$$

For any $1 \le j \le M$,

$$\begin{aligned}
&\mathbb{E}\{ d^2(Y,\omega) \mid X \in L(x,\Pi_\lambda^{(j)}) \} \\
&= \int_{L(x,\Pi_\lambda^{(j)})} \left(\int_\Omega d^2(y,\omega) \frac{\rho(y|z)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \mu(\mathrm{d}y) \right) \nu_X(\mathrm{d}z) \\
&= \int_\Omega d^2(y,\omega) \left(\int_{L(x,\Pi_\lambda^{(j)})} \frac{\rho(y|z)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \nu_X(\mathrm{d}z) \right) \mu(\mathrm{d}y).
\end{aligned}$$

By Assumption (A2), for μ-almost all $y \in \Omega$, $\forall\, \varepsilon > 0$, $\exists\, \delta_\varepsilon > 0$ (dependent on x, y) such that when $z \in \mathrm{B}(x, \|\cdot\|_2, \delta_\varepsilon)$, we have

$$|\rho(y|z) - \rho(y|x)| \le \varepsilon.$$

Thus, $\exists\ \delta_\varepsilon > 0$ such that

$$\begin{aligned}
&\left| \int_{L(x,\Pi_\lambda^{(j)})} \{\rho(y|z) - \rho(y|x)\} \nu_X(\mathrm{d}z) \right| \\
&\le \int_{L(x,\Pi_\lambda^{(j)})} |\rho(y|z) - \rho(y|x)| \, \nu_X(\mathrm{d}z) \\
&= \int_{L(x,\Pi_\lambda^{(j)}) \cap \mathrm{B}(x,\|\cdot\|_2,\delta_\varepsilon)} |\rho(y|z) - \rho(y|x)| \, \nu_X(\mathrm{d}z) \\
&\quad + \int_{L(x,\Pi_\lambda^{(j)}) \setminus \mathrm{B}(x,\|\cdot\|_2,\delta_\varepsilon)} |\rho(y|z) - \rho(y|x)| \, \nu_X(\mathrm{d}z) \\
&\le \varepsilon \, \nu_X\left(L(x,\Pi_\lambda^{(j)}) \right) \\
&\quad + 2 \sup_z \rho(y|z) \, \nu_X\left(L(x,\Pi_\lambda^{(j)}) \setminus \mathrm{B}(x, \|\cdot\|_2, \delta_\varepsilon) \right).
\end{aligned}$$

Then the following formula holds

$$\begin{aligned}
&\left| \int_{L(x,\Pi_\lambda^{(j)})} \frac{\rho(y|z)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \nu_X(\mathrm{d}z) - \rho(y|x) \right| \\
&= \frac{\left| \int_{L(x,\Pi_\lambda^{(j)})} \rho(y|z) \nu_X(\mathrm{d}z) - \int_{L(x,\Pi_\lambda^{(j)})} \rho(y|x) \nu_X(\mathrm{d}z) \right|}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \\
&\le \frac{\int_{L(x,\Pi_\lambda^{(j)})} |\rho(y|z) - \rho(y|z)| \, \nu_X(\mathrm{d}z)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \\
&\le \varepsilon + 2 \sup_z \rho(y|z) \frac{\nu_X\left(L(x,\Pi_\lambda^{(j)}) \setminus \mathrm{B}(x, \|\cdot\|_2, \delta_\varepsilon)\right)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)}.
\end{aligned}$$

Noticing that the density $f(x)$ is bounded and bounded away from zero by Assumption (A2), we further get

$$\begin{aligned}
&\frac{\nu_X\left(L(x,\Pi_\lambda^{(j)}) \setminus \mathrm{B}(x, \|\cdot\|_2, \delta_\varepsilon)\right)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \\
&\le \frac{\sup_z f(z) \cdot \mathrm{Vol}\left(L(x,\Pi_\lambda^{(j)}) \setminus \mathrm{B}(x, \|\cdot\|_2, \delta_\varepsilon)\right)}{\inf_z f(z) \cdot \mathrm{Vol}\left(L(x,\Pi_\lambda^{(j)})\right)},
\end{aligned}$$

where Vol denotes the volume of subsets in the $[0,1]^p$. By Corollary 1 of [21], $\mathrm{diam}\left(L(x,\Pi_\lambda^{(j)})\right) \to 0$ in probability when $\lambda \to \infty$. Hence, for δ_ε defined above,

$$\lim_{\lambda \to +\infty} \mathbb{P}\left\{ \mathrm{diam}\left(L(x,\Pi_\lambda^{(j)})\right) < \delta_\varepsilon \right\} = 1.$$

Obviously when $\mathrm{diam}\left(L(x,\Pi_\lambda^{(j)})\right) < \delta_\varepsilon$, we have

$$\mathrm{Vol}\left(L(x,\Pi_\lambda^{(j)}) \setminus \mathrm{B}(x, \|\cdot\|_2, \delta_\varepsilon)\right) = 0.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{ \mathrm{diam}\left(L(x,\Pi_\lambda^{(j)})\right) < \delta_\varepsilon \right\} \\
&\le \mathbb{P}\left\{ \frac{\mathrm{Vol}\left(L(x,\Pi_\lambda^{(j)}) \setminus \mathrm{B}(x, \|\cdot\|_2, \delta_\varepsilon)\right)}{\mathrm{Vol}\left(L(x,\Pi_\lambda^{(j)})\right)} = 0 \right\}.
\end{aligned}$$

Taking the limit on both sides of the above formula, we have $\frac{\mathrm{Vol}\left(L(x,\Pi_\lambda^{(j)}) \setminus \mathrm{B}(x,\|\cdot\|_2,\delta_\varepsilon)\right)}{\mathrm{Vol}\left(L(x,\Pi_\lambda^{(j)})\right)} \to 0$ in probability. Combine these arguments with $\sup_z \rho(y|z) < \infty$ due to the continuity of $\rho(y|z)$ on the compact set $[0,1]^p$, then

$$\left| \int_{L(x,\Pi_\lambda^{(j)})} \frac{\rho(y|z)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \nu_X(\mathrm{d}z) - \rho(y|x) \right| \xrightarrow{p} \varepsilon.$$

Let $\varepsilon \to 0$, we can get for μ-almost all $y \in \Omega$

$$\left| \int_{L(x,\Pi_\lambda^{(j)})} \frac{\rho(y|z)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \nu_X(\mathrm{d}z) - \rho(y|x) \right| \xrightarrow{p} 0.$$

Moreover,

$$\begin{aligned}
&\left| \int_{L(x,\Pi_\lambda^{(j)})} \frac{\rho(y|z)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} \nu_X(\mathrm{d}z) \right| \\
&\le \frac{\sup_z \rho(y|z) \cdot \nu_X\left(L(x,\Pi_\lambda^{(j)})\right)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} < \infty.
\end{aligned}$$

By the dominated convergence theorem and Assumption (A1), it follows that

$$\begin{aligned}
&\left| \mathbb{E}\left\{ d^2(Y,\omega) \mid X \in L(x,\Pi_\lambda^{(j)}) \right\} - \mathbb{E}\left\{ d^2(Y_i,\omega) \mid X = x \right\} \right| \\
&\le \int_\Omega d^2(y,\omega) \left| \int_{L(x,\Pi_\lambda^{(j)})} \frac{\rho(y|z) \nu_X(\mathrm{d}z)}{\nu_X\left(L(x,\Pi_\lambda^{(j)})\right)} - \rho(y|x) \right| \mu(\mathrm{d}y) \\
&\xrightarrow{p} 0.
\end{aligned}$$

Therefore, averaging over total M Mondrian trees, we get

$$\begin{aligned}&\tilde{F}_{x,M}(\omega)-F_x(\omega)\\&=\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}\Big\{d^2(Y,\omega)\mid X\in L(x,\Pi_\lambda^{(j)})\Big\}\\&\quad-\mathbb{E}\left\{d^2(Y_i,\omega)\mid X=x\right\}\\&\xrightarrow{p}0.\end{aligned}\tag{12}$$

Hence combine (11) and (12), it follows that for any $\omega\in\Omega$,

$$\hat{F}_{x,M}(\omega)-F_x(\omega)=o_p(1).$$

Then (ii): By Assumption (A4), each $L(x,\Pi_\lambda^{(j)}), 1\le j\le M$, will not be empty, leading to

$$\sum_{i=1}^{n}\alpha_{i,M}(x)=\frac{1}{M}\sum_{j=1}^{M}\frac{\sum_{i=1}^{n}\mathbb{1}_{\{X_i\in L(x,\Pi_\lambda^{(j)})\}}}{N(x,\Pi_\lambda^{(j)})}=1.$$

For any ω_1, $\omega_2\in\Omega$, the boundness of Ω gives

$$\begin{aligned}&\left|\hat{F}_{x,M}(\omega_1)-\hat{F}_{x,M}(\omega_2)\right|\\&\le\sum_{i=1}^{n}|\alpha_{i,M}(x)||d(Y_i,\omega_1)-d(Y_i,\omega_2)||d(Y_i,\omega_1)+d(Y_i,\omega_2)|\\&\le 2\,\mathrm{diam}(\Omega)d(\omega_1,\omega_2)\sum_{i=1}^{n}\alpha_{i,M}(x)\\&=O_p(d(\omega_1,\omega_2))\end{aligned}$$

where the O_p term is independent of ω_1 and ω_2. Hence

$$\sup_{d(\omega_1,\omega_2)<\delta}\left|\hat{F}_{x,M}(\omega_1)-\hat{F}_{x,M}(\omega_2)\right|=O_p(\delta),$$

which can deduce (ii). □

Appendix B
Proof of Theorem 2

Proof. The proof that we are about to give will be developed upon the work of [9], which focuses on local polynomial Fréchet regression under the fixed design with $X_i=i/n\in[0,1], i=1,\dots,n$. Here we consider the Fréchet Mondrian forest model under the random design, where $X_i\sim\nu_X$ on $[0,1]^p$. Let us describe it briefly. The variance inequality is initially employed to transform the error bound of the minimizer to a uniform bound of the objective function. The uniform bound is then decomposed into two distinct parts: the bias term and the variance term. By analyzing the two terms separately, the derived results combined with a peeling device can effectively bound the tail probabilities of the error. Finally, the convergence rate is obtained by integrating the tail probabilities. Following [9], define

$$\begin{aligned}\Diamond(y_1,y_2,\omega_1,\omega_2):=&d^2(y_1,\omega_1)-d^2(y_1,\omega_2)-d^2(y_2,\omega_1)\\&+d^2(y_2,\omega_2);\\\mathfrak{a}(y_1,y_2):=&\sup_{\omega_1,\omega_2\in\Omega,\omega_1\neq\omega_2}\frac{\Diamond(y_1,y_2,\omega_1,\omega_2)}{d(\omega_1,\omega_2)}.\end{aligned}$$

Define the following objective functions

$$\begin{aligned}\hat{F}_{x,M}(\omega):=&\sum_{i=1}^{n}\alpha_{i,M}(x)d^2(Y_i,\omega),\\\bar{F}_{x,M}(\omega):=&\sum_{i=1}^{n}\alpha_{i,M}(x)\mathbb{E}\big\{d^2(Y_i,\omega)\big\},\\F_x(\omega):=&\mathbb{E}\big\{d^2(Y,\omega)\big|X=x\big\},\end{aligned}$$

and

$$\begin{aligned}\hat{F}_{x,M}(\omega_1,\omega_2):=&\hat{F}_{x,M}(\omega_1)-\hat{F}_{x,M}(\omega_2),\\\bar{F}_{x,M}(\omega_1,\omega_2):=&\bar{F}_{x,M}(\omega_1)-\bar{F}_{x,M}(\omega_2),\\F_x(\omega_1,\omega_2):=&F_x(\omega_1)-F_x(\omega_2).\end{aligned}$$

Variance Inequality and Split. Using Assumption (B1) and the minimizing property of $\hat{m}_{\oplus,M}(x)$, we obtain

$$\begin{aligned}&C_{\mathsf{Vlo}}^{-1}d^2(\hat{m}_{\oplus,M}(x),m_\oplus(x))\\&\le F_x(\hat{m}_{\oplus,M}(x),m_\oplus(x))\\&\le F_x(\hat{m}_{\oplus,M}(x),m_\oplus(x))-\hat{F}_{x,M}(\hat{m}_{\oplus,M}(x),m_\oplus(x))\\&=\Big\{F_x(\hat{m}_{\oplus,M}(x),m_\oplus(x))-\bar{F}_{x,M}(\hat{m}_{\oplus,M}(x),m_\oplus(x))\Big\}+\\&\quad\Big\{\bar{F}_{x,M}(\hat{m}_{\oplus,M}(x),m_\oplus(x))-\hat{F}_{x,M}(\hat{m}_{\oplus,M}(x),m_\oplus(x))\Big\}.\end{aligned}\tag{13}$$

The first bracket represents the bias term, and the second one is the variance term.

Variance. Conditional on $\{X_i\}_{i=1}^n$ and $\Pi_{\lambda,M}$, define

$$\begin{aligned}Z_{i,x}(\omega)=&\alpha_{i,M}(x)\Big\{d^2(Y_i,\omega)-d^2(Y_i,m_\oplus(x))\Big\}\\&-\alpha_{i,M}(x)\,\mathbb{E}\Big\{d^2(Y_i,\omega)-d^2(Y_i,m_\oplus(x))\Big\}.\end{aligned}$$

Then $Z_{1,x},\dots,Z_{n,x}$ are independent and centered processes with $Z_{i,x}(m_\oplus(x))=0$. They are integrable due to Assumption (B3). By the definition of $\mathfrak{a}$,

$$\begin{aligned}&\left|Z_{i,x}(\omega_1)-Z_{i,x}(\omega_2)-Z'_{i,x}(\omega_1)+Z'_{i,x}(\omega_2)\right|\\&\le|\alpha_{i,M}(x)|\,\mathfrak{a}(Y_i,Y_i')d(\omega_1,\omega_2),\end{aligned}$$

where $Z'_{i,x}(\omega)$ and Y_i' are independent copies of $Z_{i,x}(\omega)$ and Y_i, respectively. Theorem 10 of [9] implies

$$\begin{aligned}&\mathbb{E}\left\{\sup_{\omega\in\mathrm{B}(m_\oplus(x),d,\delta)}\left|\bar{F}_{x,M}(\omega,m_\oplus(x))-\hat{F}_{x,M}(\omega,m_\oplus(x))\right|^\kappa\right\}\\&=\mathbb{E}\left\{\sup_{\omega\in\mathrm{B}(m_\oplus(x),d,\delta)}\left|\sum_{i=1}^{n}Z_{i,x}(\omega)\right|^\kappa\right\}\\&\le c_\kappa\left[\left\{\mathbb{E}\left(\sum_{i=1}^{n}\alpha_{i,M}(x)^2\mathfrak{a}(Y_i,Y_i')^2\right)^{\frac{\kappa}{2}}\right\}^{\frac{1}{\kappa}}\times\right.\\&\qquad\left.\gamma_2(\mathrm{B}(m_\oplus(x),d,\delta),d)\right]^\kappa\end{aligned}$$

for a constant c_κ depending only on κ. If $\alpha_{i,M}(x)=0$ for all $1\le i\le n$, $\mathbb{E}\left(\sum_{i=1}^{n}\alpha_{i,M}(x)^2\mathfrak{a}(Y_i,Y_i')^2\right)^{\frac{\kappa}{2}}=0$; Otherwise, let $W=\sum_{i=1}^{n}\alpha_{i,M}(x)^2$ and $v_{i,M}(x)=\alpha_{i,M}(x)^2/W$. We apply Assumption (B3) and get

$$\begin{aligned}&\mathbb{E}\left(\sum_{i=1}^{n}\alpha_{i,M}(x)^2\mathfrak{a}(Y_i,Y_i')^2\right)^{\frac{\kappa}{2}}\\&=\mathbb{E}\left(W\sum_{i=1}^{n}v_{i,M}(x)\mathfrak{a}(Y_i,Y_i')^2\right)^{\frac{\kappa}{2}}\end{aligned}$$

$$\leq W^{\frac{\kappa}{2}} \sum_{i=1}^{n} v_{i,M}(x) \mathbb{E}\mathfrak{a}(Y_i, Y_i')^{\kappa}$$
$$\leq W^{\frac{\kappa}{2}} C_{\mathsf{Mom}}^{\kappa}.$$

Hence, these two cases can be summarized as

$$\mathbb{E}\left(\sum_{i=1}^{n} \alpha_{i,M}(x)^2 \mathfrak{a}(Y_i, Y_i')^2\right)^{\frac{\kappa}{2}} \leq W^{\frac{\kappa}{2}} C_{\mathsf{Mom}}^{\kappa}.$$

By Jensen's inequality,

$$W = \sum_{i=1}^{n} \left\{ \frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{X_i \in L(x, \Pi_\lambda^{(j)})\}}}{N(x, \Pi_\lambda^{(j)})} \right\}^2$$
$$\leq \sum_{i=1}^{n} \frac{1}{M} \sum_{j=1}^{M} \left\{ \frac{\mathbb{1}_{\{X_i \in L(x, \Pi_\lambda^{(j)})\}}}{N(x, \Pi_\lambda^{(j)})} \right\}^2$$
$$\leq \frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{N(x, \Pi_\lambda^{(j)})>0\}}}{N(x, \Pi_\lambda^{(j)})}.$$

By Assumption (B2), conditional on $\{X_i\}_{i=1}^n$ and $\Pi_{\lambda,M}$,

$$\mathbb{E}\left\{ \sup_{\omega \in \mathrm{B}(m_\oplus(x), d, \delta)} \left| \bar{F}_{x,M}(\omega, m_\oplus(x)) - \hat{F}_{x,M}(\omega, m_\oplus(x)) \right|^{\kappa} \right\}$$
$$\leq c_\kappa \left\{ C_{\mathsf{Mom}} C_{\mathsf{Ent}} \max(\delta, \delta^\alpha) \left(\frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{N(x, \Pi_\lambda^{(j)})>0\}}}{N(x, \Pi_\lambda^{(j)})} \right)^{\frac{1}{2}} \right\}^{\kappa}. \tag{14}$$

Bias. Using the conditional μ-density $y \mapsto \rho(y|x)$, we can write

$$F_x(\omega, m_\oplus(x))$$
$$= \mathbb{E}\{d^2(Y, \omega) | X = x\} - \mathbb{E}\{d^2(Y, m_\oplus(x)) | X = x\}$$
$$= \int \{d^2(y, \omega) - d^2(y, m_\oplus(x))\} \rho(y|x) \mu(\mathrm{d}y).$$

Since

$$\bar{F}_{x,M}(\omega) = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N(x, \Pi_\lambda^{(j)})} \sum_{i: X_i \in L(x, \Pi_\lambda^{(j)})} \mathbb{E}\{d^2(Y_i, \omega)\},$$

it follows that

$$\bar{F}_{x,M}(\omega, m_\oplus(x))$$
$$= \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N(x, \Pi_\lambda^{(j)})} \sum_{i: X_i \in L(x, \Pi_\lambda^{(j)})} \int \{d^2(y, \omega)$$
$$- d^2(y, m_\oplus(x))\} \rho(y|X_i) \mu(\mathrm{d}y)$$
$$= \int \{d^2(y, \omega) - d^2(y, m_\oplus(x))\} \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N(x, \Pi_\lambda^{(j)})}$$
$$\sum_{i: X_i \in L(x, \Pi_\lambda^{(j)})} \rho(y|X_i) \mu(\mathrm{d}y).$$

Let $\hat{\rho}_{\lambda,M}(y|x) := \frac{1}{M} \sum_{j=1}^{M} \frac{1}{N(x, \Pi_\lambda^{(j)})} \sum_{i: X_i \in L(x, \Pi_\lambda^{(j)})} \rho(y|X_i) =: \frac{1}{M} \sum_{j=1}^{M} \hat{\rho}_\lambda^{(j)}(y|x)$ and then

$$\left| F_x(\omega, m_\oplus(x)) - \bar{F}_{x,M}(\omega, m_\oplus(x)) \right|$$
$$= \left| \int \{d^2(y, \omega) - d^2(y, m_\oplus(x))\} \{\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\} \mu(\mathrm{d}y) \right|$$
$$\leq \int \left| d^2(y, \omega) - d^2(y, m_\oplus(x)) \right| \left| \rho(y|x) - \hat{\rho}_{\lambda,M}(y|x) \right| \mu(\mathrm{d}y)$$
$$\leq d(\omega, m_\oplus(x)) \int \{d(y, \omega) + d(y, m_\oplus(x))\}$$
$$\left| \rho(y|x) - \hat{\rho}_{\lambda,M}(y|x) \right| \mu(\mathrm{d}y).$$

Recall the definition of $H(\omega_1, \omega_2)$ in Assumption (B5). By the Cauchy–Schwartz inequality,

$$\int \{d(y, \omega) + d(y, m_\oplus(x))\} \left| \rho(y|x) - \hat{\rho}_{\lambda,M}(y|x) \right| \mu(\mathrm{d}y)$$
$$\leq H(\omega, m_\oplus(x)) \left(\int \left| \rho(y|x) - \hat{\rho}_{\lambda,M}(y|x) \right|^2 \mu(\mathrm{d}y) \right)^{\frac{1}{2}}.$$

Thus,

$$\left| F_x(\hat{m}_{\oplus,M}(x), m_\oplus(x)) - \bar{F}_{x,M}(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \right|$$
$$\leq d(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \, H(\hat{m}_{\oplus,M}(x), m_\oplus(x))$$
$$\left(\int \left| \rho(y|x) - \hat{\rho}_{\lambda,M}(y|x) \right|^2 \mu(\mathrm{d}y) \right)^{\frac{1}{2}}. \tag{15}$$

Since $\mathbb{E}\{H(\hat{m}_{\oplus,M}(x), m_\oplus(x))^{\kappa} | \{X_i\}_{i=1}^n, \Pi_{\lambda,M}\}^{\frac{1}{\kappa}} \leq C_{\mathsf{Bom}}$ by (B5), (15) implies

$$\mathbb{E}\Big\{ \left| F_x(\hat{m}_{\oplus,M}(x), m_\oplus(x)) - \bar{F}_{x,M}(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \right|^{\kappa}$$
$$\mathbb{1}_{\{d(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \in [0,\delta]\}} \Big\}^{\frac{1}{\kappa}}$$
$$\leq C_{\mathsf{Bom}} \delta \left(\int \left| \rho(y|x) - \hat{\rho}_{\lambda,M}(y|x) \right|^2 \mu(\mathrm{d}y) \right)^{\frac{1}{2}}. \tag{16}$$

Peeling. For $\delta > 0$, define

$$\Delta_\delta(\omega_1, \omega_2) = \Big(\left| F_x(\omega_1, \omega_2) - \bar{F}_{x,M}(\omega_1, \omega_2) \right| + \left| \bar{F}_{x,M}(\omega_1, \omega_2) - \hat{F}_{x,M}(\omega_1, \omega_2) \right| \Big) \mathbb{1}_{\{d(\omega_1, \omega_2) \in [0,\delta]\}}.$$

Let $0 < a < b < \infty$. The inequality (13) and Markov's inequality yield

$$\mathbb{P}\Big\{ d(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \in [a, b] \Big| \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \Big\}$$
$$\leq \mathbb{P}\Big\{ a^2 \leq C_{\mathsf{Vlo}} \Delta_b(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \Big| \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \Big\}$$
$$\leq \frac{C_{\mathsf{Vlo}}^{\kappa} \mathbb{E}\Big[\{\Delta_b(\hat{m}_{\oplus,M}(x), m_\oplus(x))\}^{\kappa} \Big| \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \Big]}{a^{2\kappa}}.$$

By virtue of our previous deliberations about the variance term (14) and the bias term (16), we are able to bound the conditional expectation of $\Delta_\delta(\hat{m}_{\oplus,M}(x), m_\oplus(x))^{\kappa}$.

$$\mathbb{E}\left\{ \Delta_\delta(\hat{m}_{\oplus,M}(x), m_\oplus(x))^{\kappa} \mid \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \right\} \leq 2^{\kappa-1}$$
$$\Bigg(\mathbb{E}\Big\{ \left| F_x(\hat{m}_{\oplus,M}(x), m_\oplus(x)) - \bar{F}_{x,M}(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \right|^{\kappa} \mathbb{1}_{\{d(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \in [0,\delta]\}} \Big\} + \mathbb{E}\Big\{ \sup_{\omega \in \mathrm{B}(m_\oplus(x), d, \delta)} \left| \bar{F}_{x,M}(\omega, m_\oplus(x)) - \hat{F}_{x,M}(\omega, m_\oplus(x)) \right|^{\kappa} \Big\} \Bigg)$$

$$
\begin{aligned}
&= 2^{\kappa-1}\Bigg(\left\{ C_{\mathsf{Bom}}\delta \left(\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y) \right)^{\frac{1}{2}} \right\}^{\kappa} + \\
&\quad c_\kappa \left\{ C_{\mathsf{Mom}} C_{\mathsf{Ent}} \max(\delta, \delta^\alpha) \left(\frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})>0\}}}{N(x,\Pi_\lambda^{(j)})} \right)^{\frac{1}{2}} \right\}^{\kappa} \Bigg) \\
&\le c_\kappa \left\{ C_{\mathsf{Bom}} \left(\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y) \right)^{\frac{1}{2}} + \right. \\
&\quad \left. C_{\mathsf{Mom}} C_{\mathsf{Ent}} \left(\frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})>0\}}}{N(x,\Pi_\lambda^{(j)})} \right)^{\frac{1}{2}} \right\}^{\kappa} \times \max(\delta, \delta^\alpha)^\kappa.
\end{aligned}
$$

We are now prepared to apply peeling (also called slicing): Let $s > 0$. Set

$$
\begin{aligned}
A_{n,\lambda} =& C_{\mathsf{Vlo}} C_{\mathsf{Bom}} \left(\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y) \right)^{\frac{1}{2}} \\
&+ C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} \left(\frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})>0\}}}{N(x,\Pi_\lambda^{(j)})} \right)^{\frac{1}{2}}.
\end{aligned}
$$

It holds that

$$
\begin{aligned}
&\mathbb{P}\left\{ d(\hat{m}_{\oplus,M}(x), m_\oplus(x)) > s \mid \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \right\} \\
&\le \sum_{l=0}^{\infty} \mathbb{P}\left\{ d(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \in [2^l s, 2^{l+1} s] \middle| \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \right\} \\
&\le \sum_{l=0}^{\infty} \frac{c_\kappa A^\kappa \max\left(2^{l+1}s, (2^{l+1}s)^\alpha\right)^\kappa}{(2^l s)^{2\kappa}} \\
&\le c_\kappa A^\kappa \left(s^{-\kappa} + s^{-\kappa(2-\alpha)} \right) \sum_{l=0}^{\infty} 2^{-l\kappa(2-\alpha)} \\
&\le c_\kappa A^\kappa \left(s^{-\kappa} + s^{-\kappa(2-\alpha)} \right).
\end{aligned}
$$

We integrate the tail to bound the expectation. For this we require $\kappa > \frac{2}{2-\alpha}$. Set $B_{n,\lambda} = c_\kappa A_{n,\lambda}^\kappa$, then

$$
\begin{aligned}
&\mathbb{E}\left\{ d^2(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \mid \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \right\} \\
&= 2\int_0^\infty s\, \mathbb{P}\left\{ d(\hat{m}_{\oplus,M}(x), m_\oplus(x)) > s \middle| \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \right\} \mathrm{d}s \\
&\le 2\int_0^\infty s \min\left(1, B_{n,\lambda}\left(s^{-\kappa} + s^{-\kappa(2-\alpha)} \right)\right) \mathrm{d}s \\
&\le 2\int_0^\infty s \min\left(1, B_{n,\lambda} s^{-\kappa}\right) \mathrm{d}s + \\
&\quad 2\int_0^\infty s \min\left(1, B_{n,\lambda} s^{-\kappa(2-\alpha)}\right) \mathrm{d}s.
\end{aligned}
$$

For the first summand,

$$
\begin{aligned}
&2\int_0^\infty s \min\left(1, B_{n,\lambda} s^{-\kappa}\right) \mathrm{d}s \\
&= 2\int_0^{B_{n,\lambda}^{\frac{1}{\kappa}}} s\,\mathrm{d}s + 2B_{n,\lambda} \int_{B_{n,\lambda}^{\frac{1}{\kappa}}}^\infty s^{1-\kappa} \mathrm{d}s \\
&= B_{n,\lambda}^{\frac{2}{\kappa}} + \frac{2B_{n,\lambda}}{\kappa-2} B_{n,\lambda}^{\frac{2-\kappa}{\kappa}} \\
&= \frac{\kappa}{\kappa-2} B_{n,\lambda}^{\frac{2}{\kappa}}.
\end{aligned}
$$

Similarly,

$$
2\int_0^\infty s \min\left(1, B_{n,\lambda} s^{-\kappa(2-\alpha)}\right) \mathrm{d}s \le \frac{\kappa(2-\alpha)}{\kappa(2-\alpha)-2} B_{n,\lambda}^{\frac{2}{\kappa(2-\alpha)}}.
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\left\{ d^2(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \mid \{X_i\}_{i=1}^n, \Pi_{\lambda,M} \right\} \\
&\le c_{\alpha,\kappa} \left(A_{n,\lambda}^2 + A_{n,\lambda}^{\frac{2}{2-\alpha}} \right) \\
&\le c_{\alpha,\kappa} \left(C_{\mathsf{Vlo}} C_{\mathsf{Bom}} \right)^{\frac{2}{2-\alpha}} \left\{ \left(\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y) \right) \right. \\
&\quad \left. + \left(\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y) \right)^{\frac{1}{2-\alpha}} \right\} + \\
&\quad c_{\alpha,\kappa} \left(C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} \right)^{\frac{2}{2-\alpha}} \left(\frac{1}{M} \sum_{j=1}^{M} \frac{\mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})>0\}}}{N(x,\Pi_\lambda^{(j)})} \right) \\
&=: c_{\alpha,\kappa} \left(C_{\mathsf{Vlo}} C_{\mathsf{Bom}} \right)^{\frac{2}{2-\alpha}} T_1 + c_{\alpha,\kappa} \left(C_{\mathsf{Vlo}} C_{\mathsf{Mom}} C_{\mathsf{Ent}} \right)^{\frac{2}{2-\alpha}} T_2.
\end{aligned} \tag{17}
$$

Now we take expectation to the above terms over $\{X_i\}_{i=1}^n$ and $\Pi_{\lambda,M}$.

(i) Analysis of T_2. By Lemma 4.1 of [55],

$$
\mathbb{E}(T_2) = \mathbb{E}\left(\frac{\mathbb{1}_{\{N(x,\Pi_\lambda)>0\}}}{N(x,\Pi_\lambda)} \right) \le \mathbb{E}\left(\frac{2}{n\nu_X\left(L(x,\Pi_\lambda)\right)} \right), \tag{18}
$$

where the last inequality comes from the independence between Π_λ and $\{X_i\}_{i=1}^n$.

(ii) Analysis of T_1. By Assumption (B4),

$$
|\rho(y|x) - \rho(y|X_i)| \le L(y)\|x - X_i\|_2^\beta,
$$

the definition of $\hat{\rho}_{\lambda,M}(y|x)$ leads to

$$
\begin{aligned}
&\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y) \\
&\le \frac{1}{M} \sum_{j=1}^{M} \int \left\{ \frac{\mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})>0\}}}{N(x,\Pi_\lambda^{(j)})} \sum_{i:X_i \in L(x,\Pi_\lambda^{(j)})} |\rho(y|x) - \rho(y|X_i)| \right. \\
&\quad \left. + \mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})=0\}} \rho(y|x) \right\}^2 \mu(\mathrm{d}y) \\
&\le \frac{1}{M} \sum_{j=1}^{M} \int \left\{ L(y) \frac{\mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})>0\}}}{N(x,\Pi_\lambda^{(j)})} \sum_{i:X_i \in L(x,\Pi_\lambda^{(j)})} \|x - X_i\|_2^\beta \right. \\
&\quad \left. + \mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})=0\}} \rho(y|x) \right\}^2 \mu(\mathrm{d}y) \\
&\le \frac{1}{M} \sum_{j=1}^{M} \int \left\{ L(y) D_\lambda^{(j)}(x)^\beta + \mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})=0\}} \rho(y|x) \right\}^2 \mu(\mathrm{d}y) \\
&\le \frac{2}{M} \sum_{j=1}^{M} D_\lambda^{(j)}(x)^{2\beta} \int L^2(y) \mu(\mathrm{d}y) \\
&\quad + \frac{2}{M} \sum_{j=1}^{M} \mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})=0\}} \int \rho^2(y|x) \mu(\mathrm{d}y) \\
&\le C_{\mathsf{SmD}}^2 \frac{2}{M} \sum_{j=1}^{M} D_\lambda^{(j)}(x)^{2\beta} + C_{\mathsf{Cdl}}^2 \frac{2}{M} \sum_{j=1}^{M} \mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})=0\}},
\end{aligned}
$$

where $D_\lambda(x)$ stands for the l_2-diameter of the leaf $L(x,\Pi_\lambda^{(j)})$ containing x in the j-th Mondrian tree. Therefore,

$$\begin{aligned}
&\mathbb{E}(T_1)\\
&\le 2^{\frac{\alpha}{2-\alpha}} C_{\mathsf{SmD}}^{\frac{2}{2-\alpha}} \mathbb{E}\left(\frac{1}{M}\sum_{j=1}^{M} D_\lambda^{(j)}(x)^{2\beta} + \frac{1}{M}\sum_{j=1}^{M} D_\lambda^{(j)}(x)^{\frac{2\beta}{2-\alpha}}\right)\\
&\quad + 2^{\frac{\alpha}{2-\alpha}} C_{\mathsf{CdI}}^{\frac{2}{2-\alpha}} \mathbb{E}\left(\frac{2}{M}\sum_{j=1}^{M} \mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})=0\}}\right)\\
&\le 2^{\frac{\alpha}{2-\alpha}} C_{\mathsf{SmD}}^{\frac{2}{2-\alpha}} \left(\mathbb{E}\left\{D_\lambda(x)^{2\beta}\right\} + \mathbb{E}\left\{D_\lambda(x)^{\frac{2\beta}{2-\alpha}}\right\}\right)\\
&\quad + (2C_{\mathsf{CdI}})^{\frac{2}{2-\alpha}} \mathbb{E}\left(\mathbb{1}_{\{N(x,\Pi_\lambda)=0\}}\right)\\
&\le (2C_{\mathsf{SmD}})^{\frac{2}{2-\alpha}} p^{\frac{\beta}{2-\alpha}} \mathbb{E}\left\{\left(\frac{D_\lambda(x)}{\sqrt{p}}\right)^{2\beta}\right\}\\
&\quad + (2C_{\mathsf{CdI}})^{\frac{2}{2-\alpha}} \mathbb{E}\left(\mathbb{1}_{\{N(x,\Pi_\lambda)=0\}}\right).
\end{aligned}$$

By Corollary 1 of [21], we have

$$\mathbb{E}\left\{\left(\frac{D_\lambda(x)}{\sqrt{p}}\right)^{2\beta}\right\} \le p^{-\beta}\left(\mathbb{E}\left\{D_\lambda(x)^2\right\}\right)^{\beta} \le \left(\frac{8}{\lambda^2}\right)^{\beta}.$$

By the independence between Π_λ and $\{X_i\}_{i=1}^n$, we have

$$\begin{aligned}
\mathbb{E}\left(\mathbb{1}_{\{N(x,\Pi_\lambda)=0\}}\right) &= \mathbb{E}\left(\left\{1-\nu_X\left(L(x,\Pi_\lambda)\right)\right\}^n\right)\\
&\le \mathbb{E}\left(e^{-n\nu_X(L(x,\Pi_\lambda))}\right)\\
&\le \mathbb{E}\left(\frac{e^{-1}}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right).
\end{aligned} \tag{19}$$

Hence

$$\begin{aligned}
\mathbb{E}(T_1) \le{}& (2C_{\mathsf{SmD}})^{\frac{2}{2-\alpha}} p^{\frac{\beta}{2-\alpha}} \left(8/\lambda^2\right)^{\beta}\\
&+ (2C_{\mathsf{CdI}})^{\frac{2}{2-\alpha}} \mathbb{E}\left(\frac{e^{-1}}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right).
\end{aligned} \tag{20}$$

Finally, combining (17), (18) and (20) gives

$$\begin{aligned}
&\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(x), m_\oplus(x))\right\}\\
=&\mathbb{E}\left[\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \mid \{X_i\}_{i=1}^n, \Pi_{\lambda,M}\right\}\right]\\
\le& c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}\right)^{\frac{2}{2-\alpha}} \mathbb{E}(T_1) + c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} \mathbb{E}(T_2)\\
\le& c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{SmD}}\right)^{\frac{2}{2-\alpha}} p^{\frac{\beta}{2-\alpha}} \left(8/\lambda^2\right)^{\beta} +\\
&c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{CdI}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} \mathbb{E}\left(\frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right).
\end{aligned}$$

Integrating risk over the hypercube. Now we integrate the above inequality with respect to $x \in [0,1]^p$ to obtain the bound for the mean integrated squared error. Assume there are K_λ leaves in Π_λ, denoted by $A_1, A_2, \ldots, A_{K_\lambda}$, which is a partition of input space, then

$$\begin{aligned}
&\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X))\right\}\\
=&\int \mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(x), m_\oplus(x))\right\} \nu_X(\mathrm{d}x)\\
\le& c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{SmD}}\right)^{\frac{2}{2-\alpha}} p^{\frac{\beta}{2-\alpha}} \left(8/\lambda^2\right)^{\beta} + c_{\alpha,\kappa}(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}\\
&C_{\mathsf{CdI}}C_{\mathsf{Mom}}C_{\mathsf{Ent}})^{\frac{2}{2-\alpha}} \mathbb{E}\left(\sum_{k=1}^{K_\lambda}\int_{A_k} \frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\nu_X(\mathrm{d}x)\right)\\
\le& c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{SmD}}\right)^{\frac{2}{2-\alpha}} p^{\frac{\beta}{2-\alpha}} \left(8/\lambda^2\right)^{\beta}\\
&+ c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{CdI}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} \mathbb{E}(K_\lambda)/n\\
\le& c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{SmD}}\right)^{\frac{2}{2-\alpha}} p^{\frac{\beta}{2-\alpha}} \left(8/\lambda^2\right)^{\beta}\\
&+ c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{CdI}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} (1+\lambda)^p/n,
\end{aligned}$$

since $\mathbb{E}(K_\lambda) = (1+\lambda)^p$ by Proposition 2 of [21].

Lastly, taking $\lambda = \lambda_n \asymp n^{\frac{1}{p+2\beta}}$, we can get

$$\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X))\right\} \le O\left(n^{-\frac{2\beta}{p+2\beta}}\right).$$

□

Appendix C
Proof of Theorem 3

Proof. By the proof of Theorem 2, we have the following conclusion (see (17)):

$$\begin{aligned}
&\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(x), m_\oplus(x)) \mid \{X_i\}_{i=1}^n, \Pi_{\lambda,M}\right\}\\
\le& c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}\right)^{\frac{2}{2-\alpha}} \left\{\left(\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y)\right)\right.\\
&\left. + \left(\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y)\right)^{\frac{1}{2-\alpha}}\right\} +\\
&c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} \left(\frac{1}{M}\sum_{j=1}^{M} \frac{\mathbb{1}_{\{N(x,\Pi_\lambda^{(j)})>0\}}}{N(x,\Pi_\lambda^{(j)})}\right)\\
=:& c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}\right)^{\frac{2}{2-\alpha}} T_1 + c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} T_2.
\end{aligned} \tag{21}$$

To get a sharper upper bound, we need to make a delicate analysis of T_1. Recall that

$$\begin{aligned}
\hat{\rho}_{\lambda,M}(y|x) &= \frac{1}{M}\sum_{j=1}^{M} \hat{\rho}_\lambda^{(j)}(y|x)\\
&= \frac{1}{M}\sum_{j=1}^{M} \frac{1}{N(x,\Pi_\lambda^{(j)})} \sum_{i:X_i\in L(x,\Pi_\lambda^{(j)})} \rho(y|X_i).
\end{aligned}$$

For $1 \le j \le M$, let

$$\bar{\rho}_\lambda^{(j)}(y|x) = \mathbb{E}_X\left\{\rho(y|X) \mid X \in L(x,\Pi_\lambda^{(j)})\right\},$$

which depends on $\Pi_\lambda^{(j)}$. And let

$$\tilde{\rho}_\lambda(y|x) = \mathbb{E}\left[\bar{\rho}_\lambda^{(j)}(y|x)\right] = \mathbb{E}\left[\mathbb{E}_X\left\{\rho(y|X) \mid X \in L(x,\Pi_\lambda^{(j)})\right\}\right],$$

which is deterministic and does not depend on $\Pi_{\lambda,M}$. By Jensen's inequality and $\int \rho^2(y|x)\mu(\mathrm{d}y) \le C_{\mathsf{CdI}}^2$ for all $x \in [0,1]^p$ in Assumption (B4′),

$$\begin{aligned}
&\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y)\\
\le& \int \rho^2(y|x)\mu(\mathrm{d}y) + \int \hat{\rho}_{\lambda,M}^2(y|x)\mu(\mathrm{d}y) \le 2C_{\mathsf{CdI}}^2.
\end{aligned}$$

Taking expectation with respect to $\{X_i\}_{i=1}^n$ and $\Pi_{\lambda,M}$, we have the following decomposition based on the above result:

$$\mathbb{E}\left(\int \left|\rho(y|x) - \hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y)\right)^{\frac{1}{2-\alpha}}$$

$$
\begin{aligned}
&\le \left(2C_{\mathsf{Cdl}}^2\right)^{\frac{\alpha-1}{2-\alpha}} \mathbb{E}\left(\int \left|\rho(y|x)-\hat{\rho}_{\lambda,M}(y|x)\right|^2 \mu(\mathrm{d}y)\right) \\
&= \left(2C_{\mathsf{Cdl}}^2\right)^{\frac{\alpha-1}{2-\alpha}} \int \mathbb{E}\Big|\frac{1}{M}\sum_{j=1}^{M}\hat{\rho}_{\lambda}^{(j)}(y|x)-\rho(y|x)\Big|^2 \mu(\mathrm{d}y) \\
&\le 2\left(2C_{\mathsf{Cdl}}^2\right)^{\frac{\alpha-1}{2-\alpha}} \int \mathbb{E}\Big|\frac{1}{M}\sum_{j=1}^{M}\left(\hat{\rho}_{\lambda}^{(j)}(y|x)-\bar{\rho}_{\lambda}^{(j)}(y|x)\right)\Big|^2 \mu(\mathrm{d}y) \\
&\quad + 2\left(2C_{\mathsf{Cdl}}^2\right)^{\frac{\alpha-1}{2-\alpha}} \int \mathbb{E}\Big|\frac{1}{M}\sum_{j=1}^{M}\bar{\rho}_{\lambda}^{(j)}(y|x)-\rho(y|x)\Big|^2 \mu(\mathrm{d}y) \\
&:= 2\left(2C_{\mathsf{Cdl}}^2\right)^{\frac{\alpha-1}{2-\alpha}} T_{11} + 2\left(2C_{\mathsf{Cdl}}^2\right)^{\frac{\alpha-1}{2-\alpha}} T_{12}.
\end{aligned}
$$

(i) Analysis of T_{11}. Jensen's inequality implies that

$$
\begin{aligned}
&\mathbb{E}\Big|\frac{1}{M}\sum_{j=1}^{M}\left(\hat{\rho}_{\lambda}^{(j)}(y|x)-\bar{\rho}_{\lambda}^{(j)}(y|x)\right)\Big|^2 \\
\le&\mathbb{E}\left|\hat{\rho}_{\lambda}^{(1)}(y|x)-\bar{\rho}_{\lambda}^{(1)}(y|x)\right|^2 \\
=&\mathbb{E}\Big|\frac{1}{N(x,\Pi_\lambda)}\sum_{i:X_i\in L(x,\Pi_\lambda)}\rho(y|X_i) \\
&\quad -\mathbb{E}_X\left\{\rho(y|X)\Big|X\in L(x,\Pi_\lambda)\right\}\Big|^2 \\
=&\mathbb{E}\Big[\mathbb{E}\Big\{\Big|\frac{1}{N(x,\Pi_\lambda)}\sum_{i:X_i\in L(x,\Pi_\lambda)}\rho(y|X_i) \\
&\quad -\mathbb{E}_X\left\{\rho(y|X)\Big|X\in L(x,\Pi_\lambda)\right\}\Big|^2\Big|\Pi_\lambda, N(x,\Pi_\lambda)\Big\}\Big] \\
\le&\mathbb{E}\Big[\frac{\mathbb{1}_{\{N(x,\Pi_\lambda)>0\}}}{N(x,\Pi_\lambda)}\mathbb{E}_{X_1}\Big[\Big(\rho(y|X_1) \\
&\quad -\mathbb{E}_X\left\{\rho(y|X)\Big|X\in L(x,\Pi_\lambda)\right\}\Big)^2\Big|X_1\in L(x,\Pi_\lambda)\Big] \\
&\quad +\mathbb{1}_{\{N(x,\Pi_\lambda)=0\}}\left[\mathbb{E}_X\left\{\rho(y|X)\Big|X\in L(x,\Pi_\lambda)\right\}\right]^2\Big] \\
\le&\mathbb{E}\Big[\frac{\mathbb{1}_{\{N(x,\Pi_\lambda)>0\}}}{N(x,\Pi_\lambda)}2\cdot\mathbb{E}_X\left\{\rho^2(y|X)\Big|X\in L(x,\Pi_\lambda)\right\} \\
&\quad +\mathbb{1}_{\{N(x,\Pi_\lambda)=0\}}\mathbb{E}_X\left\{\rho^2(y|X)\Big|X\in L(x,\Pi_\lambda)\right\}\Big],
\end{aligned}
$$

where the second inequality is because $\rho(y|X_i)$ are i.i.d. for $X_i\in L(x,\Pi_\lambda)$ given Π_λ and $N(x,\Pi_\lambda)$. Then we can get

$$
\begin{aligned}
&T_{11} \\
\le&\mathbb{E}\Big[2\frac{\mathbb{1}_{\{N(x,\Pi_\lambda)>0\}}}{N(x,\Pi_\lambda)}\mathbb{E}_X\Big\{\int\rho^2(y|X)\mu(\mathrm{d}y)\Big|X\in L(x,\Pi_\lambda)\Big\} \\
&\quad +\mathbb{1}_{\{N(x,\Pi_\lambda)=0\}}\mathbb{E}_X\Big\{\int\rho^2(y|X)\mu(\mathrm{d}y)\Big|X\in L(x,\Pi_\lambda)\Big\}\Big] \\
\le&2C_{\mathsf{Cdl}}^2\mathbb{E}\left(\frac{2}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right)+C_{\mathsf{Cdl}}^2\mathbb{E}\left(\frac{e^{-1}}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right),
\end{aligned}
\tag{22}
$$

where the last inequality has utilized (18), (19) and Assumption (B4′).

(ii) Analysis of T_{12}. $\bar{\rho}_{\lambda}^{(j)}(y|x)$ are i.i.d. for $j=1,\dots,M$ with expectation $\tilde{\rho}^{(j)}(y|x)$. Therefore

$$
\begin{aligned}
&\mathbb{E}\left|\frac{1}{M}\sum_{j=1}^{M}\bar{\rho}_{\lambda}^{(j)}(y|x)-\rho(y|x)\right|^2 \\
&= \left(\tilde{\rho}_\lambda(y|x)-\rho(y|x)\right)^2+\mathrm{Var}\left(\bar{\rho}_{\lambda}^{(1)}(y|x)\right)/M.
\end{aligned}
$$

By Assumption (B4′), $\rho(y|x)$ is $L(y)$-Lipschitz w.r.t. x. Thus,

$$
\begin{aligned}
\mathrm{Var}\left(\bar{\rho}_{\lambda}^{(1)}(y|x)\right) &\le \mathbb{E}\left\{\left(\bar{\rho}_{\lambda}^{(1)}(y|x)-\rho(y|x)\right)^2\right\} \\
&\le L(y)^2\mathbb{E}\left\{D_\lambda(x)^2\right\}\le\frac{4pL(y)^2}{\lambda^2},
\end{aligned}
$$

where we have used Corollary 1 of [21]. Consequently,

$$
\begin{aligned}
&\mathbb{E}\left|\frac{1}{M}\sum_{j=1}^{M}\bar{\rho}_{\lambda}^{(j)}(y|x)-\rho(y|x)\right|^2 \\
&\le\left(\tilde{\rho}_\lambda(y|x)-\rho(y|x)\right)^2+\frac{4pL(y)^2}{\lambda^2 M}.
\end{aligned}
$$

Then

$$
T_{12}\le\int\left(\tilde{\rho}_\lambda(y|x)-\rho(y|x)\right)^2\mu(\mathrm{d}y)+C_{\mathsf{SmD}}^2\frac{4p}{\lambda^2 M}. \tag{23}
$$

By (22) and (23), we get

$$
\begin{aligned}
&\mathbb{E}\left(\int\left|\rho(y|x)-\hat{\rho}_{\lambda,M}(y|x)\right|^2\mu(\mathrm{d}y)\right)^{\frac{1}{2-\alpha}} \\
&\le c_\alpha C_{\mathsf{Cdl}}^{\frac{2}{2-\alpha}}\mathbb{E}\left(\frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right)+c_\alpha C_{\mathsf{Cdl}}^{\frac{2\alpha-2}{2-\alpha}}C_{\mathsf{SmD}}^2\frac{p}{\lambda^2 M} \\
&\quad +c_\alpha C_{\mathsf{Cdl}}^{\frac{2\alpha-2}{2-\alpha}}\int\left(\tilde{\rho}_\lambda(y|x)-\rho(y|x)\right)^2\mu(\mathrm{d}y).
\end{aligned}
\tag{24}
$$

It remains to control $\int\left(\tilde{\rho}_\lambda(y|x)-\rho(y|x)\right)^2\mu(\mathrm{d}y)$ in (24), which heavily relies on leveraging the higher smoothness of the conditional density function $\rho(y|x)$. Thanks to [21] for a thorough analysis of the Euclidean Mondrian forest, we outline a similar proof process as follows. Let us recall that $L(x,\Pi_\lambda)$ represents the leaf of Π_λ which contains $x\in[0,1]^p$ and $f(x)$ is the density of X. It is not hard to see

$$
\begin{aligned}
\tilde{\rho}_\lambda(y|x) &= \mathbb{E}\left[\frac{1}{\nu_X\left(L(x,\Pi_\lambda)\right)}\int_{[0,1]^p}\rho(y|z)f(z)\mathbb{1}_{\{z\in L(x,\Pi_\lambda)\}}\mathrm{d}z\right] \\
&= \int_{[0,1]^p}\rho(y|z)F_{f,\lambda}(x,z)\mathrm{d}z,
\end{aligned}
\tag{25}
$$

where

$$
F_{f,\lambda}(x,z)=\mathbb{E}\left[\frac{f(z)\mathbb{1}_{\{z\in L(x,\Pi_\lambda)\}}}{\nu_X\left(L(x,\Pi_\lambda)\right)}\right].
$$

In particular, $\int_{[0,1]^p}F_{f,\lambda}(x,z)\mathrm{d}z=1$ for any $x\in[0,1]^p$ (letting $\rho(y|\cdot)\equiv 1$ above). By Assumption (B4′), it holds that

$$
\begin{aligned}
&\left|\rho(y|z)-\rho(y|x)-\nabla\rho(y|x)^\top(z-x)\right| \\
&=\left|\int_0^1\left[\nabla\rho(y|x+t(z-x))-\nabla\rho(y|x)\right]^\top(z-x)\mathrm{d}t\right| \\
&\le\int_0^1 L(y)(t\|z-x\|_2)^\beta\|z-x\|_2\mathrm{d}t\le L(y)\|z-x\|_2^{1+\beta}.
\end{aligned}
$$

Now, by the triangle inequality,

$$\begin{aligned}&\left|\left|\int_{[0,1]^p}\{\rho(y|z)-\rho(y|x)\}F_{f,\lambda}(x,z)\mathrm{d}z\right|\right.\\&\left.-\left|\int_{[0,1]^p}\nabla\rho(y|x)^\top(z-x)F_{f,\lambda}(x,z)\mathrm{d}z\right|\right|\\&\leq L(y)\int_{[0,1]^p}\|z-x\|_2^{1+\beta}F_{f,\lambda}(x,z)\mathrm{d}z,\end{aligned}$$

so that, with the fact $\int_{[0,1]^p}F_{f,\lambda}(x,z)\mathrm{d}z=1$ and (25), we get

$$\begin{aligned}&|\tilde{\rho}_\lambda(y|x)-\rho(y|x)|\\&\leq\left|\nabla\rho(y|x)^\top\int_{[0,1]^p}(z-x)F_{f,\lambda}(x,z)\mathrm{d}z\right|\\&\quad+L(y)\int_{[0,1]^p}\|z-x\|_2^{1+\beta}F_{f,\lambda}(x,z)\mathrm{d}z.\end{aligned}$$

Under the assumption that the density f of X is positive and C_f-Lipschitz, the proof of Theorem 3 of [21] establishes the following bounds:

$$\begin{aligned}&\int_{[0,1]^p}\|z-x\|_2^{1+\beta}F_{f,\lambda}(x,z)\mathrm{d}z\leq\frac{f_1}{f_0}\left(\frac{2p}{\lambda^2}\right)^{(1+\beta)/2},\\&\left\|\int_{[0,1]^p}(z-x)F_{f,\lambda}(x,z)\mathrm{d}z\right\|_2^2\leq\frac{18}{\lambda^2}\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}\\&\qquad+2\left(\frac{f_1C_f}{f_0^2}\frac{3p\sqrt{p}}{\lambda^2}\right)^2,\end{aligned}$$

where $f_0=\inf_x f(x)>0$ and $f_1=\sup_x f(x)<\infty$ since f is positive and continuous. Therefore,

$$\begin{aligned}&\left(|\tilde{\rho}_\lambda(y|x)-\rho(y|x)|\right)^2\\&\leq2\left(\|\nabla\rho(y|x)\|_2^2\times\left\|\int_{[0,1]^p}(z-x)F_{f,\lambda}(x,z)\mathrm{d}z\right\|_2^2\right.\\&\qquad\left.+L(y)^2\left(\int_{[0,1]^p}\|z-x\|_2^{1+\beta}F_{f,\lambda}(x,z)\mathrm{d}z\right)^2\right)\\&\leq2L(y)^2\left[\frac{18}{\lambda^2}\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}+2\left(\frac{f_1C_f}{f_0^2}\frac{3p\sqrt{p}}{\lambda^2}\right)^2\right]\\&\quad+2L(y)^2\left(\frac{f_1}{f_0}\right)^2\left(\frac{2p}{\lambda^2}\right)^{1+\beta}\\&\leq\frac{36L(y)^2}{\lambda^2}\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}+\frac{36L(y)^2p^3}{\lambda^4}\left(\frac{f_1C_f}{f_0^2}\right)^2\\&\quad+\frac{8L(y)^2p^{1+\beta}}{\lambda^{2(1+\beta)}}\left(\frac{f_1}{f_0}\right)^2,\end{aligned}$$

which leads to

$$\begin{aligned}&\int(\tilde{\rho}_\lambda(y|x)-\rho(y|x))^2\mu(\mathrm{d}y)\\&\leq\frac{36}{\lambda^2}C_{\mathsf{SmD}}^2\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}+\frac{36p^3}{\lambda^4}\left(\frac{f_1C_f}{f_0^2}\right)^2C_{\mathsf{SmD}}^2\\&\quad+\frac{8p^{1+\beta}}{\lambda^{2(1+\beta)}}\left(\frac{f_1}{f_0}\right)^2C_{\mathsf{SmD}}^2.\end{aligned}\tag{26}$$

Plugging (26) into (24) gives us

$$\begin{aligned}&\mathbb{E}\left(\int|\rho(y|x)-\hat{\rho}_{\lambda,M}(y|x)|^2\mu(\mathrm{d}y)\right)^{\frac{1}{2-\alpha}}\leq c_\alpha C_{\mathsf{CdI}}^{\frac{2}{2-\alpha}}\times\\&\mathbb{E}\left(\frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right)+c_\alpha C_{\mathsf{CdI}}^{\frac{2\alpha-2}{2-\alpha}}C_{\mathsf{SmD}}^2\left\{\frac{p}{\lambda^2M}+\frac{1}{\lambda^2}\right.\\&\left.\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}+\left(\frac{f_1C_f}{f_0^2}\right)^2\frac{p^3}{\lambda^4}+\left(\frac{f_1}{f_0}\right)^2\frac{p^{1+\beta}}{\lambda^{2(1+\beta)}}\right\}.\end{aligned}\tag{27}$$

Naturally, taking $\alpha=1$, we have

$$\begin{aligned}&\mathbb{E}\left(\int|\rho(y|x)-\hat{\rho}_{\lambda,M}(y|x)|^2\mu(\mathrm{d}y)\right)\leq c_\alpha C_{\mathsf{CdI}}^2\times\\&\mathbb{E}\left(\frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right)+c_\alpha C_{\mathsf{SmD}}^2\left\{\frac{p}{\lambda^2M}+\frac{1}{\lambda^2}\right.\\&\left.\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}+\left(\frac{f_1C_f}{f_0^2}\right)^2\frac{p^3}{\lambda^4}+\left(\frac{f_1}{f_0}\right)^2\frac{p^{1+\beta}}{\lambda^{2(1+\beta)}}\right\}.\end{aligned}\tag{28}$$

Finally, combine (21), (18), (27) and (28), then we get

$$\begin{aligned}&\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(x),m_\oplus(x))\right\}\\&=\mathbb{E}\left[\mathbb{E}\left\{d^2(\hat{m}_{\oplus,M}(x),m_\oplus(x))\mid\{X_i\}_{i=1}^n,\Pi_{\lambda,M}\right\}\right]\\&\leq c_{\alpha,\kappa}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}C_{\mathsf{CdI}}C_{\mathsf{Mom}}C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}}\mathbb{E}\left(\frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right)+\\&c_{\alpha,\kappa}C_{\mathsf{SmD}}^2C_{\mathsf{CdI}}^{\frac{2\alpha-2}{2-\alpha}}\left(C_{\mathsf{Vlo}}C_{\mathsf{Bom}}\right)^{\frac{2}{2-\alpha}}\left\{\frac{p}{\lambda^2M}+\left(\frac{f_1C_f}{f_0^2}\right)^2\frac{p^3}{\lambda^4}+\right.\\&\left.\left(\frac{f_1}{f_0}\right)^2\frac{p^{1+\beta}}{\lambda^{2(1+\beta)}}+\frac{1}{\lambda^2}\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}\right\}.\end{aligned}\tag{29}$$

Integrating risk over the hypercube. Now consider the hypercube $B_\varepsilon=[\varepsilon,1-\varepsilon]^p$ with $\varepsilon\in(0,1/2)$. We integrate $\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}$ over X conditionally on $X\in B_\varepsilon$, specifically,

$$\begin{aligned}&\mathbb{E}\left\{\sum_{j=1}^p e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}\middle|X\in B_\varepsilon\right\}\\&=\sum_{j=1}^p\mathbb{E}\left\{e^{-\lambda\left[x_{(j)}\wedge(1-x_{(j)})\right]}\middle|X\in B_\varepsilon\right\}\\&\leq\frac{pf_1}{f_0(1-2\varepsilon)^p}\int_\varepsilon^{1-\varepsilon}e^{-\lambda[u\wedge(1-u)]}\mathrm{d}u\\&=\frac{pf_1}{f_0(1-2\varepsilon)^p}\times2\int_\varepsilon^{1/2}e^{-\lambda u}\mathrm{d}u\\&\leq\frac{e^{-\lambda\varepsilon}}{\lambda}\frac{2pf_1}{f_0(1-2\varepsilon)^p},\end{aligned}\tag{30}$$

due to $f_0\leq f(x)\leq f_1$ for any $x\in[0,1]^p$. In addition, by Proposition 2 of [21], we have

$$\begin{aligned}&\int\mathbb{E}\left(\frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\right)\nu_X(\mathrm{d}x)\\&=\mathbb{E}\left(\sum_{k=1}^{K_\lambda}\int_{A_k}\frac{1}{n\nu_X\left(L(x,\Pi_\lambda)\right)}\nu_X(\mathrm{d}x)\right)\end{aligned}$$

$$= \frac{\mathbb{E}(K_\lambda)}{n} \le \frac{(1+\lambda)^p}{n},$$

where $A_1, A_2, \ldots, A_{K_\lambda}$ are K_λ leaves in Π_λ and form a partition of input space. Hence, integrating $1/\{n\nu_X(L(x,\Pi_\lambda))\}$ over X conditionally on $X \in B_\varepsilon$ gives

$$\begin{aligned}
&\mathbb{E}\left\{ \frac{1}{n\nu_X\big(L(X,\Pi_\lambda)\big)} \middle| X \in B_\varepsilon \right\} \\
&\le \mathbb{P}\big(X \in B_\varepsilon\big)^{-1} \int \mathbb{E}\left(\frac{1}{n\nu_X\big(L(x,\Pi_\lambda)\big)} \right) \nu_X(\mathrm{d}x) \\
&\le \frac{1}{f_0(1-2\varepsilon)^p} \frac{(1+\lambda)^p}{n}. \qquad (31)
\end{aligned}$$

With the conditional expectation results (30) and (31), the upper bound of integrating (29) over X conditionally on $X \in B_\varepsilon$ follows as

$$\begin{aligned}
&\mathbb{E}\left\{ d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X)) \mid X \in B_\varepsilon \right\} \\
&\le c_{\alpha,\kappa} \left(C_{\mathsf{Vlo}} C_{\mathsf{Bom}} C_{\mathsf{Cdl}} C_{\mathsf{Mom}} C_{\mathsf{Ent}}\right)^{\frac{2}{2-\alpha}} \frac{1}{f_0(1-2\varepsilon)^p} \frac{(1+\lambda)^p}{n} + \\
&c_{\alpha,\kappa} C_{\mathsf{SmD}}^2 C_{\mathsf{Cdl}}^{\frac{2\alpha-2}{2-\alpha}} \left(C_{\mathsf{Vlo}} C_{\mathsf{Bom}}\right)^{\frac{2}{2-\alpha}} \left\{ \frac{p}{\lambda^2 M} + \left(\frac{f_1 C_f}{f_0^2}\right)^2 \frac{p^3}{\lambda^4} + \right. \\
&\left. \left(\frac{f_1}{f_0}\right)^2 \frac{p^{1+\beta}}{\lambda^{2(1+\beta)}} + \frac{f_1}{f_0(1-2\varepsilon)^p} \frac{p e^{-\lambda\varepsilon}}{\lambda^3} \right\}. \qquad (32)
\end{aligned}$$

In particular, for a fixed $\varepsilon \in (0, 1/2)$, (32) writes

$$\begin{aligned}
&\mathbb{E}\left\{ d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X)) \mid X \in B_\varepsilon \right\} \\
&= O\left(\frac{\lambda^p}{n} + \frac{1}{\lambda^{2(1+\beta)}} + \frac{1}{\lambda^2 M} \right).
\end{aligned}$$

By optimizing the terms on the right-hand side, one can choose $\lambda = \lambda_n \asymp n^{1/(p+2s)}$ and $M = M_n \gtrsim \lambda_n^{2\beta} \asymp n^{2\beta/(p+2s)}$, where $s = 1+\beta \in (1,2]$. This yields the rate $O\left(n^{-2s/(p+2s)}\right)$.

In contrast, when $\varepsilon = 0$, we have $e^{-\lambda\varepsilon} = 1$, so that the inequality (32) becomes in this case

$$\begin{aligned}
&\mathbb{E}\left\{ d^2(\hat{m}_{\oplus,M}(X), m_\oplus(X)) \right\} \\
&= O\left(\frac{\lambda^p}{n} + \frac{1}{\lambda^{3\wedge 2(1+\beta)}} + \frac{1}{\lambda^2 M} \right).
\end{aligned}$$

For $2s \le 3$ (equivalent to $\beta \le 1/2$), this gives the same rate as before with identical parameter choices. Conversely, when $2s > 3$, the rate becomes suboptimal at $O\left(n^{-3/(p+3)}\right)$, obtained by setting $M_n \gtrsim \lambda_n \asymp n^{1/(p+3)}$. This completes the proof of all statements in Theorem 3. □

Appendix D
Proof of Proposition 1

We prove Proposition 1 by analyzing Examples 1–4 separately.

A. Proof of Example 1

Proof. Since $\mathcal{S}_2^+$ with the log-Euclidean metric is a Hadamard space, by Remark 1, Assumption (B1) holds with $C_{\mathsf{Vlo}} = 1$ and Assumption (B5) holds with $C_{\mathsf{Bom}} = c_\kappa C_{\mathsf{Mom}} C_{\mathsf{Len}} C_{\mathsf{Int}}$, where the unknown constants will be determined progressively in the subsequent discussion. For any $\mathcal{B} \subset \mathcal{S}_2^+$, let $\log(\mathcal{B}) = \{\log(S) : S \in \mathcal{B}\}$. Since $d(S_1, S_2) = \|\log(S_1) - \log(S_2)\|_{\mathrm{F}}$, we can regard $\log(\mathcal{B})$ as a subset of $\mathcal{R}^4$ endowed with the standard Euclidean metric $\|\cdot\|_2$, then

$$\begin{aligned}
N(\mathcal{B}, d, r) &= N(\log(\mathcal{B}), \|\cdot\|_2, r) \\
&\le N(\mathrm{B}_2(\mathrm{diam}(\log(\mathcal{B}), \|\cdot\|_2)/2), \|\cdot\|_2, r) \\
&\le N(\mathrm{B}_2(\mathrm{diam}(\mathcal{B}, d)/2), \|\cdot\|_2, r) \\
&\le \left(\frac{\mathrm{diam}(\mathcal{B}, d)}{r} + 1 \right)^4,
\end{aligned}$$

where $\mathrm{B}_2(b)$ is a Euclidean ball with radius $b \in \mathcal{R}$ in $\mathcal{R}^4$ and the last inequlity comes from Corollary 4.2.13 of [56]. To simplify the bound a bit, in the non-trivial range $r \in (0, \mathrm{diam}(\mathcal{B}, d)]$, we have

$$N(\mathcal{B}, d, r) \le \left(\frac{2\,\mathrm{diam}(\mathcal{B}, d)}{r} \right)^4;$$

in the trivial range where $r > \mathrm{diam}(\mathcal{B}, d)$, $\mathcal{B}$ can be covered by just one r-ball in $\mathcal{S}_2^+$, so $N(\mathcal{B}, d, r) = 1$. Thus

$$\begin{aligned}
\gamma_2(\mathcal{B}, d) &\le c \int_0^\infty \sqrt{\log(N(\mathcal{B}, d, r))}\mathrm{d}r \\
&\le 2c \int_0^{\mathrm{diam}(\mathcal{B},d)} \sqrt{\log\left(\frac{2\,\mathrm{diam}(\mathcal{B}, d)}{r} \right)}\mathrm{d}r \\
&\le 4c \cdot \mathrm{diam}(\mathcal{B}, d).
\end{aligned}$$

That is to asy, Assumption (B2) holds with $\alpha = 1$ and $C_{\mathsf{Ent}} = \max\{1, 4c\}$. By the spectral decomposition,

$$Y = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} e^{1+f(X,\varepsilon)} & 0 \\ 0 & e^{1+f(X,\varepsilon)} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Therefore, we have

$$\begin{aligned}
&\log(Y) \\
&= \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1+f(X,\varepsilon) & 0 \\ 0 & 1+f(X,\varepsilon) \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & f(X,\varepsilon) \\ f(X,\varepsilon) & 1 \end{pmatrix}.
\end{aligned}$$

Based on it, the Fréchet regression function in Example 1 is

$$\begin{aligned}
m_\oplus(x) &= \operatorname*{argmin}_{\omega \in \mathcal{S}_2^+} \mathbb{E}\left\{ \|\log(Y) - \log(\omega)\|_{\mathrm{F}}^2 \mid X = x \right\} \\
&= \exp(q_x)
\end{aligned}$$

with

$$q_x = \begin{pmatrix} 1 & \mathbb{E}f(x,\varepsilon) \\ \mathbb{E}f(x,\varepsilon) & 1 \end{pmatrix} = \begin{pmatrix} 1 & x+1 \\ x+1 & 1 \end{pmatrix}.$$

Then for all $\kappa > 2$ and $x \in [0,1]$,

$$\begin{aligned}
&\mathbb{E}\left\{ d^\kappa(Y, m_\oplus(x)) \mid X = x \right\} \\
&= \mathbb{E}\left\{ \left\| \begin{pmatrix} 1 & f(x,\varepsilon) \\ f(x,\varepsilon) & 1 \end{pmatrix} - \begin{pmatrix} 1 & x+1 \\ x+1 & 1 \end{pmatrix} \right\|_{\mathrm{F}}^\kappa \right\} \\
&= \mathbb{E}\left\{ (2\varepsilon^2)^{\kappa/2} \right\} \\
&= 2^\kappa \sigma^\kappa \frac{\Gamma(\frac{\kappa+1}{2})}{\sqrt{\pi}}.
\end{aligned}$$

Thus (B3) holds with $C_{\mathsf{Mom}} = \max\{1, 2\sigma(\Gamma(\frac{\kappa+1}{2})/\sqrt{\pi})^{1/\kappa}\}$. As for Assumption (B4′),

$$\begin{aligned}
&\sup_{x_1,x_2\in[0,1]} d(m_\oplus(x_1), m_\oplus(x_2)) \\
=&\sup_{x_1,x_2\in[0,1]} \|q_{x_1} - q_{x_2}\|_{\mathrm{F}} \\
=&\sup_{x_1,x_2\in[0,1]} \left\| \begin{pmatrix} 0 & x_1 - x_2 \\ x_1 - x_2 & 0 \end{pmatrix} \right\|_{\mathrm{F}} \\
\le& \sqrt{2},
\end{aligned}$$

so we can take $C_{\mathsf{Len}} = \sqrt{2}$. Let the probability measure μ on $\mathcal{S}_2^+$ be the distribution of

$$\exp\left(\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix} \right), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

For any $x_0 \in [0,1]$, it holds that

$$\begin{aligned}
&\int d^2(y, m_\oplus(x_0))\mu(\mathrm{d}y) \\
=&\int \|\log(y) - \log(m_\oplus(x_0))\|_{\mathrm{F}}^2 \mu(\mathrm{d}y) \\
=&\int \left\| \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix} - \begin{pmatrix} 1 & x_0+1 \\ x_0+1 & 1 \end{pmatrix} \right\|_{\mathrm{F}}^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}} \mathrm{d}\varepsilon \\
=&2\sigma^2 + 2(x_0+1)^2.
\end{aligned}$$

Thus C_{Int} can be $2\sigma^2 + 2(x_0+1)^2$ with $x_0 \in [0,1]$. Additioanlly, conditional on $X = x$, the μ-density of Y is

$$\rho(y|x) = e^{-\frac{(x+1)^2 - 2(x+1)(\log(y))_{12}}{2\sigma^2}},$$

where $(\log(y))_{12}$ is the $(1,2)$-th entry of the matrix $\log(y)$. The derivatives of $\rho(y|x)$ with respect to x are

$$\begin{aligned}
\nabla\rho(y|x) =& -\frac{x+1-(\log(y))_{12}}{\sigma^2} e^{-\frac{(x+1)^2 - 2(x+1)(\log(y))_{12}}{2\sigma^2}}, \\
\nabla^2\rho(y|x) =& \left\{ \left(\frac{x+1-(\log(y))_{12}}{\sigma^2} \right)^2 - \frac{1}{\sigma^2} \right\} \times \\
& e^{-\frac{(x+1)^2 - 2(x+1)(\log(y))_{12}}{2\sigma^2}}.
\end{aligned}$$

Since $\nabla^2\rho(y|x)$ is bounded with respect to x due to its continuity on $[0,1]$, $\nabla\rho(y|x)$ is Lipschitz continuous with respect to x, leading to $\beta = 1$. The rest of Assumption (B4′) can be easily checked.

Lastly, we derive the lower bound for the prediction error of a single Fréchet Mondrian tree estimator. Given the i.i.d training sample $\{(X_i, Y_i)\}_{i=1}^n$ from (X, Y), the prediction given by a Fréchet Mondrian tree can be written as

$$\begin{aligned}
\hat{m}_\oplus(x) &= \operatorname*{argmin}_{\omega\in\mathcal{S}_2^+} \frac{1}{N(x,\Pi_\lambda)} \sum_{i:X_i\in L(x,\Pi_\lambda)} d^2(Y_i, \omega) \\
&= \operatorname*{argmin}_{\omega\in\mathcal{S}_2^+} \frac{1}{N(x,\Pi_\lambda)} \sum_{i:X_i\in L(x,\Pi_\lambda)} \|\log(Y_i) - \log(\omega)\|_{\mathrm{F}}^2 \\
&= \exp(\hat{q}_x)
\end{aligned}$$

with

$$(\hat{q}_x)_{ij} = \begin{cases} 1, & \text{if } i = j, \\ \frac{1}{N(x,\Pi_\lambda)} \sum_{i:X_i\in L(x,\Pi_\lambda)} (X_i + 1 + \varepsilon_i), & \text{if } i \neq j. \end{cases}$$

Let $Z_i = X_i + 1 + \varepsilon_i$ and

$$\hat{m}(x) = \frac{1}{N(x,\Pi_\lambda)} \sum_{i:X_i\in L(x,\Pi_\lambda)} Z_i,$$

which is just the prediction given by a Euclidean Mondrian tree based on $\{X_i, Z_i\}_{i=1}^n$. Notably, when the leaf $L(x,\Pi_\lambda)$ is empty, adopting the convention $\hat{m}_\oplus(x) = e \cdot I_2$ is exactly equivalent to setting $\hat{m}(x) = 0$. Through the above analysis, the Fréchet regression problem in Example 1, where the response variable is symmetric positive-definite matrices, can be equivalently reformulated as the following Euclidean regression problem

$$Z = m(X) + \varepsilon = X + 1 + \varepsilon.$$

This equivalence allows us to indirectly obtain the lower bound of the risk for Mondrian tree regression on $\mathcal{S}_2^+$ by applying the result of [21]. Specifically, by Proposition 3 of [21], there exists an absolute constant C_0 such that

$$\inf_{\lambda\in\mathcal{R}_+} \mathbb{E}\big\{\hat{m}(X) - m(X)\big\}^2 \ge C_0 \wedge \frac{1}{4}\left(\frac{3\sigma^2}{n}\right)^{2/3}$$

when $n \ge 18$. Since

$$\hat{m}_\oplus(x) = \exp\left(\begin{pmatrix} 1 & \hat{m}(x) \\ \hat{m}(x) & 1 \end{pmatrix} \right)$$

and

$$m_\oplus(x) = \exp\left(\begin{pmatrix} 1 & m(x) \\ m(x) & 1 \end{pmatrix} \right),$$

the mean integrated squared error of the Fréchet Mondrian tree estimator $\hat{m}_\oplus(x)$ is

$$\begin{aligned}
&\inf_{\lambda\in\mathcal{R}_+} \mathbb{E}\left\{d^2(\hat{m}_\oplus(X), m_\oplus(X))\right\} \\
=&\inf_{\lambda\in\mathcal{R}_+} \mathbb{E}\left\{\|\log(\hat{m}_\oplus(X)) - \log(m_\oplus(X))\|_{\mathrm{F}}^2\right\} \\
=&2\inf_{\lambda\in\mathcal{R}_+} \mathbb{E}\big\{\hat{m}(X) - m(X)\big\}^2 \\
\ge& C_1 \wedge \frac{1}{2}\left(\frac{3\sigma^2}{n}\right)^{2/3}.
\end{aligned}$$

□

B. *Proof of Example 2*

Proof. The metric induced by the inner product in $L^2([0,1])$ is

$$d(f,g) = \langle f \ominus g, f \ominus g\rangle^{1/2} = \left[\int_0^1 \left\{f(t) - g(t)\right\}^2 \mathrm{d}t\right]^{1/2}.$$

Although $L^2([0,1])$ is an infinite-dimensional space, it would change nothing if we restrict the response space from $L^2([0,1])$ to its subset $\mathcal{H} = \{f_0 \oplus (c \odot g_0) : c \in \mathcal{R}\}$ for the setting in Example 2. Now we take $\Omega = \mathcal{H}$. Then the Fréchet regression function in Example 2 can be expressed by

$$\begin{aligned}
m_\oplus(x) &= \operatorname*{argmin}_{\omega\in\mathcal{H}} \mathbb{E}\left[\int_0^1 \left\{Y(t) - \omega(t)\right\}^2 \mathrm{d}t \mid X = x\right] \\
&= \mathbb{E}\left\{Y \mid X = x\right\} \\
&= f_0 \oplus \big(x \odot g_0\big),
\end{aligned} \tag{33}$$

where $\mathbb{E}$ in $\mathbb{E}\{Y|X=x\}$ refers to the Bochner integral. Therefore, for Assumption (B1),

$$\begin{aligned}
&F_x(\omega) - F_x(m_\oplus(x)) \\
&= \mathbb{E}\left\{d^2(Y,\omega) \mid X=x\right\} - \mathbb{E}\left\{d^2(Y, m_\oplus(x)) \mid X=x\right\} \\
&= \mathbb{E}\left[\int_0^1 \left\{(x+\varepsilon-c)\cdot g_0(t)\right\}^2 \mathrm{d}t\right] - \mathbb{E}\left[\int_0^1 \left\{\varepsilon\cdot g_0(t)\right\}^2 \mathrm{d}t\right] \\
&= \int_0^1 \left\{(c-x)\cdot g_0(t)\right\}^2 \mathrm{d}t \\
&= d^2(\omega, m_\oplus(x)),
\end{aligned}$$

which implies $C_{\mathsf{Vlo}} = 1$. We continue to check Assumption (B2). For any $c_1, c_2 \in \mathcal{R}$, it holds that

$$\begin{aligned}
&d(f_0 \oplus (c_1 \odot g_0), f_0 \oplus (c_2 \odot g_0)) \\
&= \left[\int_0^1 \{(c_2-c_2)\cdot g_0(t)\}^2 \mathrm{d}t\right]^{1/2} \\
&= \|g_0\| \cdot |c_1 - c_2|.
\end{aligned} \tag{34}$$

For any $\mathcal{B} \subset \mathcal{H}$, let

$$\mathcal{C} = \{c \in \mathcal{R} : f_0 \oplus (c \odot g_0) \in \mathcal{B}\} \subset \mathcal{R}.$$

By (34), we have

$$\begin{aligned}
N(\mathcal{B}, d, r) &= N(\mathcal{C}, |\cdot|, r/\|g_0\|) \\
&\le N(\mathrm{B}_2(\mathrm{diam}(\mathcal{C}, |\cdot|)/2), |\cdot|, r/\|g_0\|) \\
&\le N(\mathrm{B}_2(\mathrm{diam}(\mathcal{B}, d)/2\|g_0\|), |\cdot|, r/\|g_0\|) \\
&\le \frac{\mathrm{diam}(\mathcal{B}, d)}{r} + 1,
\end{aligned}$$

where $\mathrm{B}_2(b)$ is a Euclidean ball with radius $b \in \mathcal{R}$ in $\mathcal{R}$ and the last inequlity comes from Corollary 4.2.13 of [56]. Following similar arguments to the proof of Example 1, it can be proved that Assumption (B2) holds with $\alpha = 1$. Since

$$\begin{aligned}
\mathbb{E}\big\{d^\kappa(Y, m_\oplus(x))\big|X=x\big\} &= \mathbb{E}\left\{\int_0^1 \varepsilon^2 \cdot g_0(t)^2 \mathrm{d}t\right\}^{\kappa/2} \\
&= \|g_0\|^\kappa \cdot \mathbb{E}\big\{(\varepsilon^2)^{\kappa/2}\big\} \\
&= 2^{\kappa/2}\sigma^\kappa \|g_0\|^\kappa \frac{\Gamma(\frac{\kappa+1}{2})}{\sqrt{\pi}}
\end{aligned}$$

for all $\kappa > 2$ and $x \in [0,1]$, we can take $C_{\mathsf{Mom}} = \max\{1, \sqrt{2}\sigma\|g_0\|(\Gamma(\frac{\kappa+1}{2})/\sqrt{\pi})^{1/\kappa}\}$ in the assumption (B3). As for Assumption (B4′),

$$\begin{aligned}
\sup_{x_1,x_2\in[0,1]} d(m_\oplus(x_1), m_\oplus(x_2)) &= \sup_{x_1,x_2\in[0,1]} |x_1 - x_2| \cdot \|g_0\| \\
&\le \|g_0\|,
\end{aligned}$$

so we can take $C_{\mathsf{Len}} = \max\{1, \|g_0\|\}$. Let the probability measure μ on $\mathcal{H}$ be the distribution of

$$f_0 \oplus \big(\varepsilon \odot g_0\big), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

For any $x_0 \in [0,1]$, it holds that

$$\begin{aligned}
&\int d^2(y, m_\oplus(x_0))\mu(\mathrm{d}y) \\
&= \|g_0\|^2 \cdot \int (\varepsilon - x_0)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}} \mathrm{d}\varepsilon \\
&= \big(\sigma^2 + x_0^2\big) \cdot \|g_0\|^2.
\end{aligned}$$

Thus C_{Int} can be $\max\{1, \big(\sigma^2 + x_0^2\big) \cdot \|g_0\|^2\}$ with $x_0 \in [0,1]$. Additioanlly, conditional on $X = x$, the μ-density of Y is

$$\rho(y|x) = e^{-\frac{x^2 - 2x\{(Y \ominus f_0)/g_0\}}{2\sigma^2}}.$$

By similar arguments to the proof of Example 1, $\nabla\rho(y|x)$ is Lipschitz continuous with respect to x, leading to $\beta = 1$. The rest of Assumption (B4′) can be easily checked. There remains the assumption (B5) to be verified. Given the i.i.d training sample $\{(X_i, Y_i)\}_{i=1}^n$ from (X, Y), the prediction given by a Fréchet Mondrian tree can be written as

$$\begin{aligned}
\hat{m}_\oplus(x) &= \operatorname*{argmin}_{\omega\in\mathcal{H}} \frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} d^2(Y_i, \omega) \\
&= \frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} f_0 \oplus \big\{\big(X_i + \varepsilon_i\big) \odot g_0\big\}.
\end{aligned} \tag{35}$$

By the expression of $m_\oplus(x)$ in (33) and $\hat{m}_\oplus(x)$ in (35), we take

$$c_1 = \frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} (X_i + \varepsilon_i), \quad c_2 = x,$$

and let $y = f_0 \oplus (c_y \odot g_0)$ following the measure μ, then

$$\begin{aligned}
&H(\hat{m}_\oplus(x), m_\oplus(x)) \\
&= \left\{\int \big(|c_y - c_1| \cdot \|g_0\| + |c_y - c_2| \cdot \|g_0\|\big)^2 \mu(\mathrm{d}y)\right\}^{\frac{1}{2}} \\
&= \|g_0\| \cdot \left\{\int \big(|c_y - c_1| + |c_y - c_2|\big)^2 \mu(\mathrm{d}y)\right\}^{\frac{1}{2}}.
\end{aligned}$$

The triangle inequality tells

$$\begin{aligned}
&\int \big(|c_y - c_1| + |c_y - c_2|\big)^2 \mu(\mathrm{d}y) \\
&\le \int \big(|c_1 - c_2| + 2|c_y - c_2|\big)^2 \mu(\mathrm{d}y) \\
&\le 2|c_1 - c_2|^2 + 4\int |\varepsilon - x|^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}} \mathrm{d}\varepsilon \\
&\le 2|c_1 - c_2|^2 + 4\big(\sigma^2 + x^2\big).
\end{aligned}$$

Given $\{X_i\}_{i=1}^n, \Pi_\lambda$, we have

$$\begin{aligned}
&\mathbb{E}\big\{H(\hat{m}_\oplus(x), m_\oplus(x))^\kappa \big| \{X_i\}_{i=1}^n, \Pi_\lambda\big\} \\
&\le \|g_0\|^\kappa \cdot \mathbb{E}\left\{2|c_1 - c_2|^2 + 4\big(\sigma^2 + x^2\big)\right\}^{\kappa/2} \\
&\le c_\kappa \|g_0\|^\kappa \cdot \left\{\mathbb{E}\big(|c_1 - c_2|^\kappa\big) + \sigma^\kappa + x^\kappa\right\},
\end{aligned}$$

where

$$\mathbb{E}\left(|c_1 - c_2|^\kappa\right) = \mathbb{E}\left|\frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} (X_i + \varepsilon_i) - x\right|^\kappa$$

is bounded by x^κ if $N(x, \Pi_\lambda) = 0$; otherwise, it is bounded by $2^{\kappa-1}\left\{\frac{1}{N(x,\Pi_\lambda)} \sum_{i: X_i \in L(x,\Pi_\lambda)} |X_i - x|^\kappa + 2^{\kappa/2}\sigma^\kappa \frac{\Gamma(\frac{\kappa+1}{2})}{\sqrt{\pi}}\right\}$. Thus

$$\mathbb{E}\big\{H(\hat{m}_\oplus(x), m_\oplus(x))^\kappa \big| \{X_i\}_{i=1}^n, \Pi_\lambda\big\} \le c_\kappa \|g_0\|^\kappa \cdot (\sigma^\kappa + 1)$$

and we can choose $C_{\mathsf{Bom}} = \max\{1, c_\kappa\|g_0\| \cdot (\sigma + 1)\}$ with some constant c_κ.

Lastly, we check the lower bound for the prediction error of a single Fréchet Mondrian tree estimator. From (35), we further have

$$\hat{m}_{\oplus}(x) = (f_0 \ominus g_0) \oplus \left(\left\{ \frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} (X_i + 1 + \varepsilon_i) \right\} \odot g_0 \right).$$

Let $Z_i = X_i + 1 + \varepsilon_i$ and

$$\hat{m}(x) = \frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} Z_i,$$

which is just the prediction given by a Euclidean Mondrian tree based on $\{X_i, Z_i\}_{i=1}^n$. Notably, when the leaf $L(x, \Pi_\lambda)$ is empty, adopting the convention $f_0 \ominus g_0$ is exactly equivalent to setting $\hat{m}(x) = 0$. Noticing that

$$m_{\oplus}(x) = f_0 \oplus (x \odot g_0) = (f_0 \ominus g_0) \oplus (\{x + 1\} \odot g_0),$$

the Fréchet regression problem in Example 2, where the response variable is functional data, can be equivalently reformulated as the following Euclidean regression problem

$$Z = m(X) + \varepsilon = X + 1 + \varepsilon.$$

This equivalence allows us to indirectly obtain the lower bound of the risk for Mondrian tree regression on $L^2([0,1])$ by applying the results of [21]. Specifically, by Proposition 3 of [21], there exists an absolute constant C_0 such that

$$\inf_{\lambda \in \mathcal{R}_+} \mathbb{E}\{\hat{m}(X) - m(X)\}^2 \geq C_0 \wedge \frac{1}{4} \left(\frac{3\sigma^2}{n} \right)^{2/3}$$

when $n \geq 18$. Since

$$\hat{m}_{\oplus}(x) = (f_0 \ominus g_0) \oplus (\hat{m}(x) \odot g_0),$$
$$m_{\oplus}(x) = (f_0 \ominus g_0) \oplus (m(x) \odot g_0),$$

the mean integrated squared error of the Fréchet Mondrian tree estimator $\hat{m}_{\oplus}(x)$ is

$$\begin{aligned} &\inf_{\lambda \in \mathcal{R}_+} \mathbb{E}\left\{ d^2(\hat{m}_{\oplus}(X), m_{\oplus}(X)) \right\} \\ &= \int_0^1 g_0(t)^2 \mathrm{d}t \cdot \inf_{\lambda \in \mathcal{R}_+} \mathbb{E}\{\hat{m}(X) - m(X)\}^2 \\ &\geq C_1 \wedge \frac{\|g_0\|^2}{4} \left(\frac{3\sigma^2}{n} \right)^{2/3}. \end{aligned}$$

□

C. Proof of Example 3

Proof. Since $\mathcal{U}^2$ with the Aitchison metric is a Hilbert space and is thus a Hadamard space, by Remark 1, Assumption (B1) holds with $C_{\mathsf{Vlo}} = 1$ and Assumption (B5) holds with $C_{\mathsf{Bom}} = c_\kappa C_{\mathsf{Mom}} C_{\mathsf{Len}} C_{\mathsf{Int}}$, where the unknown constants will be determined progressively in the subsequent discussion. For any $\mathcal{B} \subset \mathcal{U}^2$, consider the following transformation

$$\widetilde{\mathcal{B}} = \left\{ g(u) = \left(\log \frac{u_{(1)}}{h(u)}, \log \frac{u_{(2)}}{h(u)}, \log \frac{u_{(3)}}{h(u)} \right)^\top : u \in \mathcal{B} \right\} \tag{36}$$

with $h(u) = \left(\prod_{i=1}^3 u_{(i)} \right)^{1/3}$. Noting that $d(u, v) = \|g(u) - g(v)\|_2$ for any $u, v \in \mathcal{U}^2$, if we consider $\widetilde{\mathcal{B}}$ as a subset of $\mathcal{R}^3$ endowed with the standard Euclidean metric $\|\cdot\|_2$, then

$$N(\mathcal{B}, d, r) = N(\widetilde{\mathcal{B}}, \|\cdot\|_2, r) \leq \left(\frac{\mathrm{diam}(\mathcal{B}, d)}{r} + 1 \right)^3,$$

where $\mathrm{B}_2(b)$ is a Euclidean ball with radius b in $\mathcal{R}^3$ and the last inequlity comes from Corollary 4.2.13 of [56]. Following an analysis similar to the proof of Example 1, it can be proved that the assumption (B2) holds with $\alpha = 1$. Since

$$Y = \left(\frac{1}{f(X, \varepsilon_1, \varepsilon_2)}, \frac{e^{X+1+\varepsilon_1}}{f(X, \varepsilon_1, \varepsilon_2)}, \frac{e^{2X+2+\varepsilon_2}}{f(X, \varepsilon_1, \varepsilon_2)} \right)^\top,$$

by the same transformation as in (36), we get

$$g(Y) = \left(-X - 1 - \frac{1}{3}\varepsilon_1 - \frac{1}{3}\varepsilon_2, \frac{2}{3}\varepsilon_1 - \frac{1}{3}\varepsilon_2, X + 1 - \frac{1}{3}\varepsilon_1 + \frac{2}{3}\varepsilon_2 \right)^\top.$$

Let $m(x) = \mathbb{E}\{g(Y) \mid X = x\}$, then

$$m(x) = (-x - 1, 0, x + 1)^\top.$$

Further, the Fréchet regression function in Example 3 can be calculated by

$$m_{\oplus}(x) = \operatorname*{argmin}_{\omega \in \mathcal{U}^2} \mathbb{E}\left\{ \|g(Y) - g(\omega)\|_2^2 \mid X = x \right\} = g^{-1}(m(x)),$$

where g^{-1} is the inverse function of g:

$$g^{-1}(a) = \left(\frac{e^{a_{(1)}}}{\sum_{i=1}^3 e^{a_{(i)}}}, \frac{e^{a_{(2)}}}{\sum_{i=1}^3 e^{a_{(i)}}}, \frac{e^{a_{(3)}}}{\sum_{i=1}^3 e^{a_{(i)}}} \right)^\top$$

with $a = \left(a_{(1)}, a_{(2)}, a_{(3)}\right)^\top$. Then for all $\kappa > 2$ and $x \in [0,1]$,

$$\begin{aligned} &\mathbb{E}\{d^\kappa(Y, m_{\oplus}(x)) \mid X = x\} \\ &= \mathbb{E}\left\{ \|g(Y) - g(m_{\oplus}(x))\|_2^\kappa \mid X = x \right\} \\ &= \mathbb{E}\left\{ \left(\frac{2}{3}\varepsilon_1^2 + \frac{2}{3}\varepsilon_2^2 - \frac{2}{3}\varepsilon_1\varepsilon_2 \right)^{\kappa/2} \right\} \\ &\leq 2^\kappa \sigma^\kappa \frac{\Gamma(\frac{\kappa+1}{2})}{\sqrt{\pi}}. \end{aligned}$$

Thus (B3) holds with $C_{\mathsf{Mom}} = \max\{1, 2\sigma(\Gamma(\frac{\kappa+1}{2})/\sqrt{\pi})^{1/\kappa}\}$. As for Assumption (B4′),

$$\begin{aligned} &\sup_{x_1, x_2 \in [0,1]} d(m_{\oplus}(x_1), m_{\oplus}(x_2)) \\ &= \sup_{x_1, x_2 \in [0,1]} \|m(x_1) - m(x_2)\|_2 \leq \sqrt{2}, \end{aligned}$$

so we can take $C_{\mathsf{Len}} = \sqrt{2}$. Let the probability measure μ on $\mathcal{U}^2$ be the distribution of

$$\left(\frac{1}{1 + e^{\varepsilon_1} + e^{\varepsilon_2}}, \frac{e^{\varepsilon_1}}{1 + e^{\varepsilon_1} + e^{\varepsilon_2}}, \frac{e^{\varepsilon_2}}{1 + e^{\varepsilon_1} + e^{\varepsilon_2}} \right)^\top,$$

where $\varepsilon_1, \varepsilon_1 \sim \mathcal{N}(0, \sigma^2)$ are independent. For any $x_0 \in [0,1]$, it holds that

$$\int d^2(y, m_{\oplus}(x_0)) \mu(\mathrm{d}y) = \int \|g(y) - m(x_0)\|_2^2 \mu(\mathrm{d}y)$$

$$=\frac{4}{3}\sigma^2+2(x_0+1)^2.$$

Thus C_{Int} can be $4\sigma^2/3+2(x_0+1)^2$ with $x_0\in[0,1]$. Additioanlly, conditional on $X=x$, the μ-density of Y is

$$\rho(y|x)=e^{-\frac{5(x+1)^2-(x+1)\left(2\log(y_{(2)})+4\log(y_{(3)})-6\log(y_{(1)})\right)}{2\sigma^2}},$$

where $y_{(i)}$ is the i-th element of y. By similar arguments to the proof of Example 1, $\nabla\rho(y|x)$ is Lipschitz continuous with respect to x, leading to $\beta=1$. The rest of the assumption (B4′) can be easily checked.

Lastly, we check the lower bound for the prediction error of a single Fréchet Mondrian tree estimator. Given the i.i.d training sample $\{(X_i,Y_i)\}_{i=1}^n$ from (X,Y), the prediction given by a Fréchet Mondrian tree can be written as

$$\begin{aligned}\hat{m}_{\oplus}(x)&=\operatorname*{argmin}_{\omega\in\mathcal{S}_2^+}\frac{1}{N(x,\Pi_\lambda)}\sum_{i:X_i\in L(x,\Pi_\lambda)}d^2\left(Y_i,\omega\right)\\&=g^{-1}(\hat{m}(x))\end{aligned}$$

with

$$\hat{m}(x)=\frac{1}{N(x,\Pi_\lambda)}\sum_{i:X_i\in L(x,\Pi_\lambda)}Z_i,$$

where $Z_i=g(Y_i)=(-X_i-1-\frac{1}{3}\varepsilon_{i1}-\frac{1}{3}\varepsilon_{i2},\frac{2}{3}\varepsilon_{i1}-\frac{1}{3}\varepsilon_{i2},X_i+1-\frac{1}{3}\varepsilon_{i1}+\frac{2}{3}\varepsilon_{i2})^\top$ and $\varepsilon_{i1},\varepsilon_{i2}$ are realizations of the noise variables $\varepsilon_1,\varepsilon_2$ in the i-th observation. It is clear that $\hat{m}(x)$ corresponds to the predictions given by a Euclidean Mondrian tree based on $\{X_i,Z_i\}_{i=1}^n$. Notably, when the leaf $L(x,\Pi_\lambda)$ is empty, adopting the convention $\hat{m}_{\oplus}(x)=(1/3,1/3,1/3)^\top$ is exactly equivalent to setting $\hat{m}(x)=0$. Through the above analysis, the Fréchet regression problem in Example 3, where the response variable is compositional data, can be equivalently reformulated as the Euclidean regression problem $Z=(m_{(1)}(X)-\frac{1}{3}\varepsilon_1-\frac{1}{3}\varepsilon_2,m_{(2)}(X)+\frac{2}{3}\varepsilon_1-\frac{1}{3}\varepsilon_2,m_{(3)}(X)-\frac{1}{3}\varepsilon_1+\frac{2}{3}\varepsilon_2)^\top=(-X-1-\frac{1}{3}\varepsilon_1-\frac{1}{3}\varepsilon_2,\frac{2}{3}\varepsilon_1-\frac{1}{3}\varepsilon_2,X+1-\frac{1}{3}\varepsilon_1+\frac{2}{3}\varepsilon_2)^\top$. This equivalence allows us to indirectly obtain the lower bound of the risk for Mondrian tree regression on $\mathcal{U}^2$ by applying the results of [21]. Specifically, by Proposition 3 of [21], there exists an absolute constant C_0 such that

$$\inf_{\lambda\in\mathcal{R}_+}\mathbb{E}\big\{\hat{m}_{(1)}(X)-m_{(2)}(X)\big\}^2\geq C_0\wedge\frac{1}{4}\left(\frac{2\sigma^2}{3n}\right)^{2/3},$$

$$\inf_{\lambda\in\mathcal{R}_+}\mathbb{E}\big\{\hat{m}_{(3)}(X)-m_{(3)}(X)\big\}^2\geq C_0\wedge\frac{1}{4}\left(\frac{5\sigma^2}{3n}\right)^{2/3},$$

when $n\geq 18$. Since

$$\hat{m}_{\oplus}(X)=g^{-1}(\hat{m}(x)),\quad m_{\oplus}(X)=g^{-1}(m(x)),$$

the mean integrated squared error of the Fréchet Mondrian tree estimator $\hat{m}_{\oplus}(x)$ is

$$\begin{aligned}&\inf_{\lambda\in\mathcal{R}_+}\mathbb{E}\left\{d^2(\hat{m}_{\oplus}(X),m_{\oplus}(X))\right\}\\&=\inf_{\lambda\in\mathcal{R}_+}\mathbb{E}\left\{\|\hat{m}(x)-m(x)\|_2^2\right\}\\&\geq\inf_{\lambda\in\mathcal{R}_+}\mathbb{E}\big\{\big(\hat{m}_{(1)}(X)-m_{(1)}(X)\big)^2+\big(\hat{m}_{(3)}(X)-m_{(3)}(X)\big)^2\big\}\\&\geq C_1\wedge\frac{1}{2}\left(\frac{2\sigma^2}{3n}\right)^{2/3}.\end{aligned}$$

□

D. Proof of Example 4

Proof. Since $\mathcal{W}_2(\mathcal{R})$ is a Hadamard space, by Remark 1, Assumption (B1) holds with $C_{\mathsf{Vlo}}=1$ and Assumption (B5) holds with $C_{\mathsf{Bom}}=c_\kappa C_{\mathsf{Mom}}C_{\mathsf{Len}}C_{\mathsf{Int}}$, where the unknown constants will be determined progressively in the subsequent discussion. Let $\omega\in\mathcal{W}_2(\mathcal{R})$ be an arbitrary probability distribution, and denote by $Q(\omega)$ its quantile function. Let d_{L_2} denote the standard L_2 metric on the space $L_2([0,1])$. Inspired by Proposition 1 in [2], we present the following argument to verify the assumption (B1). For any $\mathcal{B}\subset\mathcal{W}_2(\mathcal{R})$, the definition of Wasserstein distance yields

$$\begin{aligned}&N\left(\mathcal{B},d,r\right)\leq N\left(Q(\mathcal{B}),d_{L_2},r/2\right)\\&\leq N\Big(\mathrm{B}_{\mathrm{diam}(Q(\mathcal{B}),d_{L_2})}(Q(\omega_0))\cap Q(\mathcal{W}_2(\mathcal{R})),d_{L_2},r/2\Big)\end{aligned}$$

where $\mathrm{B}_{\mathrm{diam}(Q(\mathcal{B}),d_{L_2})}(Q(\omega_0))$ refers to the L_2 ball of radius $\mathrm{diam}(Q(\mathcal{B}),d_{L_2})$ centered at $Q(\omega_0)$ with some $\omega_0\in\mathcal{B}$. Let $\{g_u\}_{u\in U}\subset L_2([0,1])$ be a minimal r-cover of $\mathrm{B}_1(Q(\omega_0))\cap Q(\mathcal{W}_2(\mathcal{R}))$. For $\delta>0$ Define $\tilde{g}_u=Q+\delta\left(g_u-Q\right)$, then the collection $\{\tilde{g}_u\}_{u\in U}$ forms a δr-cover of $\mathrm{B}_\delta(Q(\omega_0))\cap Q(\mathcal{W}_2(\mathcal{R}))$. Thus

$$\begin{aligned}&N\left(\mathcal{B},d,r\right)\\&\leq N\Big(\mathrm{B}_{\mathrm{diam}(Q(\mathcal{B}),d_{L_2})}(Q(\omega_0))\cap Q(\mathcal{W}_2(\mathcal{R})),d_{L_2},r/2\Big)\\&\leq N\Big(\mathrm{B}_1(Q(\omega_0))\cap Q(\mathcal{W}_2(\mathcal{R})),d_{L_2},r/2\,\mathrm{diam}(Q(\mathcal{B}),d_{L_2})\Big)\\&\leq\exp\left(C\cdot\mathrm{diam}(Q(\mathcal{B}),d_{L_2})/r\right)\\&=\exp\left(C\cdot\mathrm{diam}(\mathcal{B},d)/r\right).\end{aligned}$$

where the last inequality is due to Theorem 2.7.5 of [34] and C is independent of r. Then

$$\gamma_2(\mathcal{B},d)\leq c\int_0^\infty\sqrt{\log(N(\mathcal{B},d,r))}\mathrm{d}r\leq 2c\sqrt{C}\,\mathrm{diam}(\mathcal{B},d).$$

That is to asy, the assumption (B2) can hold with $\alpha=1$ and $C_{\mathsf{Ent}}=\max\{1,2c\sqrt{C}\}$.

Next, we verify the remaining assumptions. Since the Fréchet mean of normal distributions under the Wasserstein distance remains Gaussian, it would change nothing if we restrict the response space from $\mathcal{W}^2(\mathcal{R})$ to its subset $\mathcal{G}=\{\mathcal{N}(a,1):a\in\mathcal{R}\}$ when calculating the (sample) Fréchet mean. From now on, we take the optimization range Ω to be $\mathcal{G}$. For any $\mathcal{N}(a_1,1),\mathcal{N}(a_2,1)\in\mathcal{G}$, [57] gives

$$d\left(\mathcal{N}\left(a_1,1\right),\mathcal{N}\left(a_2,1\right)\right)=\left|a_1-a_2\right|.$$

Then Fréchet regression function in Example 4 is

$$m_{\oplus}(x)=\operatorname*{argmin}_{\omega\in\mathcal{G}}\mathbb{E}\left\{d^2\left(Y,\omega\right)\mid X=x\right\}=\mathcal{N}\left(x+1,1\right).$$

It follows that

$$\mathbb{E}\big\{d^\kappa(Y,m_{\oplus}(x))\big|X=x\big\}=\mathbb{E}\left\{|\varepsilon|^\kappa\right\}=2^{\kappa/2}\sigma^\kappa\frac{\Gamma(\frac{\kappa+1}{2})}{\sqrt{\pi}}$$

for all $\kappa>2$ and $x\in[0,1]$. Then we can choose $C_{\mathsf{Mom}}=\max\{1,\sqrt{2}\sigma(\Gamma(\frac{\kappa+1}{2})/\sqrt{\pi})^{1/\kappa}\}$ in Assumption (B3). As for Assumption (B4′),

$$\sup_{x_1,x_2\in[0,1]}d(m_{\oplus}(x_1),m_{\oplus}(x_2))=\sup_{x_1,x_2\in[0,1]}|x_1-x_2|\leq 1,$$

so we can take $C_{\mathsf{Len}} = 1$. Let the probability measure μ on $\mathcal{W}_2(\mathcal{R})$ be the distribution of

$$\mathcal{N}(\varepsilon, 1), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

For any $x_0 \in [0,1]$, it holds that

$$\begin{aligned}\int d^2(y, m_\oplus(x_0))\mu(\mathrm{d}y) &= \int (\varepsilon - x_0 - 1)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}} \mathrm{d}\varepsilon \\ &= \sigma^2 + (x_0+1)^2.\end{aligned}$$

Thus C_{Int} can be $\sigma^2 + (x_0+1)^2$ with $x_0 \in [0,1]$. Additioanlly, conditional on $X = x$, the μ-density of Y is

$$\rho(y|x) = e^{-\frac{(x+1)^2 - 2(x+1)a_y}{2\sigma^2}},$$

where a_y is the location parameter of the normal distribution $y = \mathcal{N}(a_y, 1)$. Similar to the proof of Example 1, $\nabla\rho(y|x)$ is Lipschitz continuous with respect to x, leading to $\beta = 1$. The rest of Assumption (B4′) can be easily checked.

Lastly, we check the lower bound for the prediction error of a single Fréchet Mondrian tree estimator. Given the i.i.d training sample $\{(X_i, Y_i)\}_{i=1}^n$ from (X, Y), the prediction given by a Fréchet Mondrian tree can be written as

$$\begin{aligned}\hat{m}_\oplus(x) =& \operatorname*{argmin}_{\omega \in \mathcal{G}} \frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} d^2(Y_i, \omega) \\ =& \mathcal{N}\left(\frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} \big(X_i + 1 + \varepsilon_i\big), 1\right).\end{aligned}$$

Let $Z_i = X_i + 1 + \varepsilon_i$ and

$$\hat{m}(x) = \frac{1}{N(x, \Pi_\lambda)} \sum_{i: X_i \in L(x, \Pi_\lambda)} Z_i,$$

which is just the prediction given by a Euclidean Mondrian tree based on $\{X_i, Z_i\}_{i=1}^n$. Notably, when the leaf $L(x, \Pi_\lambda)$ is empty, adopting the convention $\mathcal{N}(0,1)$ is exactly equivalent to setting $\hat{m}(x) = 0$. Through the above analysis, the Fréchet regression problem in Example 4, where the response variable is probability distributions, can be equivalently reformulated as the following Euclidean regression problem

$$Z = m(X) + \varepsilon = X + 1 + \varepsilon.$$

This equivalence allows us to indirectly obtain the lower bound of the risk for Mondrian tree regression on $\mathcal{W}^2(\mathcal{R})$ by applying the results of [21]. Specifically, by Proposition 3 of [21], there exists an absolute constant C_0 such that

$$\inf_{\lambda \in \mathcal{R}_+} \mathbb{E}\big\{\hat{m}(X) - m(X)\big\}^2 \geq C_0 \wedge \frac{1}{4}\left(\frac{3\sigma^2}{n}\right)^{2/3}$$

when $n \geq 18$. Since

$$\hat{m}_\oplus(X) = \mathcal{N}\left(\hat{m}(X), 1\right), \quad m_\oplus(x) = \mathcal{N}\left(m(X), 1\right),$$

the mean integrated squared error of the Fréchet Mondrian tree estimator $\hat{m}_\oplus(x)$ is

$$\begin{aligned}\inf_{\lambda \in \mathcal{R}_+} \mathbb{E}\left\{d^2(\hat{m}_\oplus(X), m_\oplus(X))\right\} = \inf_{\lambda \in \mathcal{R}_+} \mathbb{E}\big\{\hat{m}(X) - m(X)\big\}^2 \\ \geq C_0 \wedge \frac{1}{4}\left(\frac{3\sigma^2}{n}\right)^{2/3}.\end{aligned}$$

□

REFERENCES

[1] M. Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," in *Annales de l'institut Henri Poincaré*, vol. 10, no. 4, 1948, pp. 215–310.

[2] A. Petersen and H.-G. Müller, "Fréchet regression for random objects with euclidean predictors," *The Annals of Statistics*, vol. 47, no. 2, pp. 691–719, 2019.

[3] S. Bhattacharjee and H.-G. Müller, "Concurrent object regression," *Electronic Journal of Statistics*, vol. 16, no. 2, pp. 4031–4089, 2022.

[4] ——, "Geodesic mixed effects models for repeatedly observed/longitudinal random objects," *Journal of the American Statistical Association*, pp. 1–23, 2025.

[5] ——, "Single index Fréchet regression," *The Annals of Statistics*, vol. 51, no. 4, pp. 1770–1798, 2023.

[6] A. Ghosal, W. Meiring, and A. Petersen, "Fréchet single index models for object response regression," *Electronic Journal of Statistics*, vol. 17, no. 1, pp. 1074–1112, 2023.

[7] S. Bhattacharjee, B. Li, and L. Xue, "Nonlinear global Fréchet regression for random objects via weak conditional expectation," *The Annals of Statistics*, vol. 53, no. 1, pp. 117–143, 2025.

[8] Z. Lin and H.-G. Müller, "Total variation regularized Fréchet regression for metric-space valued data," *The Annals of Statistics*, vol. 49, no. 6, pp. 3510–3533, 2021.

[9] C. Schötz, "Nonparametric regression in nonstandard spaces," *Electronic Journal of Statistics*, vol. 16, no. 2, pp. 4679–4741, 2022.

[10] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[11] G. Biau, L. Devroye, and G. Lugosi, "Consistency of random forests and other averaging classifiers." *Journal of Machine Learning Research*, vol. 9, no. 9, pp. 2015–2033, 2008.

[12] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1063–1095, 2012.

[13] E. Scornet, G. Biau, and J.-P. Vert, "Consistency of random forests," *The Annals of Statistics*, vol. 43, no. 4, pp. 1716–1741, 2015.

[14] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[15] J. Klusowski, "Sharp analysis of a simple model for random forests," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, 2021, pp. 757–765.

[16] C.-M. Chi, P. Vossler, Y. Fan, and J. Lv, "Asymptotic properties of high-dimensional random forests," *The Annals of Statistics*, vol. 50, no. 6, pp. 3415–3438, 2022.

[17] W. Peng, T. Coleman, and L. Mentch, "Rates of convergence for random forests via generalized U-statistics," *Electronic Journal of Statistics*, vol. 16, no. 1, pp. 232–292, 2022.

[18] J. M. Klusowski and P. M. Tian, "Large scale prediction with decision trees," *Journal of the American Statistical Association*, vol. 119, no. 545, pp. 525–537, 2024.

[19] B. Lakshminarayanan, D. M. Roy, and Y. W. Teh, "Mondrian forests: Efficient online random forests," *Advances in Neural Information Processing Systems*, vol. 27, pp. 3140–3148, 2014.

[20] J. Mourtada, S. Gaïffas, and E. Scornet, "Amf: Aggregated mondrian forests for online learning," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 83, no. 3, pp. 505–533, 2021.

[21] J. Mourtada, S. Gaïffas, and E. Scornet, "Minimax optimal rates for mondrian trees and forests," *The Annals of Statistics*, vol. 48, no. 4, pp. 2253–2276, 2020.

[22] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *The Annals of Statistics*, vol. 10, no. 4, pp. 1040–1053, 1982.

[23] M. D. Cattaneo, J. M. Klusowski, and W. G. Underwood, "Inference with Mondrian random forests," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, p. qkaf077, 2025.

[24] E. O'Reilly and N. M. Tran, "Minimax rates for high-dimensional random tessellation forests," *Journal of Machine Learning Research*, vol. 25, no. 89, pp. 1–32, 2024.

[25] C. Ying and Z. Yu, "Fréchet sufficient dimension reduction for random objects," *Biometrika*, vol. 109, no. 4, pp. 975–992, 2022.

[26] Q. Zhang, L. Xue, and B. Li, "Dimension reduction for Fréchet regression," *Journal of the American Statistical Association*, vol. 119, no. 548, pp. 2733–2747, 2024.

[27] L. Capitaine, J. Bigot, R. Thiébaut, and R. Genuer, "Fréchet random forests for metric space valued regression with non euclidean predictors," *Journal of Machine Learning Research*, vol. 25, no. 355, pp. 1–41, 2024.

[28] R. Qiu, Z. Yu, and R. Zhu, "Random forest weighted local fréchet regression with random objects," *Journal of Machine Learning Research*, vol. 25, no. 107, pp. 1–69, 2024.

[29] P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," *Journal of Machine Learning Research*, vol. 18, no. 181, pp. 1–18, 2018.

[30] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 9, no. 3, p. e1301, 2019.

[31] L. Mentch and S. Zhou, "Randomization as regularization: A degrees of freedom explanation for random forest success," *Journal of Machine Learning Research*, vol. 21, no. 171, pp. 1–36, 2020.

[32] A. Curth, A. Jeffares, and M. van der Schaar, "Why do random forests work? understanding tree ensembles as self-regularizing adaptive smoothers," *arXiv preprint arXiv:2402.01502*, 2024.

[33] C. E. Ginestet, A. Simmons, and E. D. Kolaczyk, "Weighted Fréchet means as convex combinations in metric spaces: properties and generalized median inequalities," *Statistics & Probability Letters*, vol. 82, no. 10, pp. 1859–1863, 2012.

[34] A. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

[35] K.-T. Sturm, "Probability measures on metric spaces of nonpositive curvature," *In Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces (Paris, 2002). Contemp. Math.*, vol. 338, pp. 357–390, 2003.

[36] B. Afsari, "Riemannian l^p center of mass: existence, uniqueness, and convexity," *Proceedings of the American Mathematical Society*, vol. 139, no. 2, pp. 655–673, 2011.

[37] N. Meinshausen and G. Ridgeway, "Quantile regression forests." *Journal of Machine Learning Research*, vol. 7, no. 6, 2006.

[38] A. Bloniarz, A. Talwalkar, B. Yu, and C. Wu, "Supervised neighborhoods for distributed nonparametric regression," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 1450–1459.

[39] M. Talagrand, *Upper and lower bounds for stochastic processes*. Springer, 2014, vol. 60.

[40] D. Burago, Y. Burago, S. Ivanov *et al.*, *A course in metric geometry*. American Mathematical Society Providence, 2001, vol. 33.

[41] W. Härdle, *Applied nonparametric regression*. Cambridge university press, 1990, no. 19.

[42] L. Györfi, M. Kohler, A. Krzyzak, H. Walk *et al.*, *A distribution-free theory of nonparametric regression*. Springer, 2002, vol. 1.

[43] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal of Matrix Analysis and Applications*, vol. 29, no. 1, pp. 328–347, 2007.

[44] Z. Lin, "Riemannian geometry of symmetric positive definite matrices via cholesky decomposition," *SIAM Journal on Matrix Analysis and Applications*, vol. 40, no. 4, pp. 1353–1370, 2019.

[45] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 3, pp. 735–747, 2005.

[46] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian framework for tensor computing," *International Journal of computer vision*, vol. 66, no. 1, pp. 41–66, 2006.

[47] T. Hsing and R. Eubank, *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

[48] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, "Functional data analysis," *Annual Review of Statistics and its application*, vol. 3, no. 1, pp. 257–295, 2016.

[49] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[50] M. Greenacre, "Compositional data analysis," *Annual Review of Statistics and its Application*, vol. 8, no. 1, pp. 271–299, 2021.

[51] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[52] Y. Chen, Z. Lin, and H.-G. Müller, "Wasserstein regression," *Journal of the American Statistical Association*, vol. 118, no. 542, pp. 869–882, 2023.

[53] B. Kloeckner, "A geometric study of Wasserstein spaces: Euclidean spaces," *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, vol. 9, no. 2, pp. 297–323, 2010.

[54] Y. Chen, A. Gajardo, J. Fan, Q. Zhong, P. Dubey, K. Han, S. Bhattacharjee, and H.-G. Müller, "frechet: Statistical analysis for random objects and non-Euclidean data," *R package version 0.2.0.*, 2020.

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

[56] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[57] M. Gelbrich, "On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces," *Mathematische Nachrichten*, vol. 147, no. 1, pp. 185–203, 1990.

Rui Qiu received the B.Sc. degree in mathematics from Central China Normal University, Wuhan, China, in 2018, and the Ph.D. degree in statistics from East China Normal University, Shanghai, China, in 2024.

He is currently a postdoctoral researcher with the School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing, China. His main research has been published in the *Annals of Statistics*, *Journal of the Royal Statistical Society, Series B*, and *Journal of Machine Learning Research*. His current research interests include non-Euclidean data analysis and statistical machine learning.

Fang Yao received the B.Sc. degree in statistics from University of Science and Technology of China, Hefei, China, in 2000, and the Ph.D. degree in statistics from University of California, Davis, CA, USA, in 2003.

He is currently a Chair Professor with the School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing, China. He has published more than 40 research articles in the top-tier journals of statistics and related interdisciplinary fields, including the *Annals of Statistics*, *Journal of the Royal Statistical Society, Series B*, *Biometrika*, *Journal of the American Statistical Association*, *IEEE Transactions on Signal Processing*, and *Journal of Machine Learning Research*. His current research interests include complex-structured data analysis, machine/deep learning, and statistical modeling of partial/ordinary differential equations.

He is a Fellow of the Institute of Mathematical Statistics (IMS) and the American Statistical Association (ASA), and an elected member of the International Statistical Institute (ISI). He has served as the Editor for the *Canadian Journal of Statistics* (2019–2021), and served on editorial boards for a number of statistical journals, including the *Annals of Statistics*, *Journal of the American Statistical Association*, and *Journal of Computational and Graphical Statistics*.

Zhou Yu received the B.Sc. degree from Hefei University of Technology, Hefei, China, in 2004, and the Ph.D. degree from East China Normal University, Shanghai, China, in 2010.

He is currently a Professor and Ph.D. Supervisor with the School of Statistics, KLATASDS-MOE, East China Normal University, Shanghai, China. He has published more than 50 research articles, including the *Annals of Statistics*, *Journal of the Royal Statistical Society, Series B*, *Biometrika*, *Journal of the American Statistical Association*, *IEEE Transactions on Information Theory*, and *Journal of Machine Learning Research*. His current research interests include statistical analysis of high-dimensional data and statistical machine learning.