EXACT FINITE-SAMPLE VARIANCE DECOMPOSITION OF SUBAGGING: A SPECTRAL FILTERING PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard resampling ratios, such as the bootstrap-inspired $\alpha \approx 0.632$, are widely used in ensemble learning, yet their finite-sample interaction with the functional complexity of the base learner remains poorly understood. We use the Hoeffding–ANOVA decomposition to derive the first exact finite-sample variance decomposition for subagging with any symmetric square-integrable base learner, where n denotes the training sample size and k the subsample size. The resulting identity shows that subagging acts as a deterministic spectral filter: the c-th order interaction component is attenuated by an explicit combinatorial factor $\gamma_c(n, k)$, bounded by and converging to α^c when the subsampling ratio $\alpha = k/n$ is held fixed. This explains why fixed resampling ratios can under-regularize high-capacity interpolators, whose variance is concentrated in high-order interactions, and why smaller α can suppress such noise exponentially. Building on this mechanism, we characterize how the optimal subsampling ratio shifts with the learner's interaction complexity and propose Complexity-Guided Adaptive Subsampling (CGAS), which empirically improves generalization over static resampling baselines.

1 INTRODUCTION

Ensemble methods, especially Bagging (Breiman, 1996a) and Subagging, remain foundational because they reduce the prediction variance of unstable base learners without inflating bias (Breiman, 2001b; Sagi & Rokach, 2018; Zhou, 2021). In practice, these methods often use fixed resampling ratios: bootstrap sampling contains about 63.2% unique training instances per base learner in the large-sample limit (Breiman, 1996b; Efron & Tibshirani, 1997), while subagging commonly adopts heuristics such as $\alpha = 0.5$ or 0.8 (Bühlmann & Yu, 2002).

Classical theory explains ensemble variance reduction through asymptotic smoothing (Bühlmann & Yu, 2002), treating the base learner's variance as an aggregate quantity. This view is insufficient for modern high-capacity learners, such as deep unpruned decision trees, which can interpolate data through high-order multi-sample interactions (Wyner et al., 2017; Belkin et al., 2019; Zhang et al., 2016). For such models, variance is not uniform across functional components but can be concentrated in high-order terms.

The resulting gap is both practical and theoretical. Practitioners tune the subsampling ratio α mostly by search (Martínez-Muñoz & Suárez, 2010), while existing theory does not give a deterministic rule linking the optimal ratio to the learner's functional complexity. Mathematically, a subagged symmetric learner is an incomplete, infinite-order U-statistic (Hoeffding, 1992; Lee, 2019). Direct Hoeffding–ANOVA variance expansions condition on macroscopic subsample overlaps, producing hypergeometric sums across interaction orders. Consequently, foundational works (Mentch & Hooker, 2016; Wager & Athey, 2018) and recent advances in variance estimation (Xu et al., 2024) resort to asymptotic limits. These limits collapse high-order structural dependencies into first-order Hájek projections, completely bypassing finite-sample combinatorics.

While effective for statistical inference, this asymptotic simplification is inadequate for analyzing modern interpolators. By projecting variance onto first-order main effects, the asymptotic limit eliminates the high-order structural dependencies where interpolation noise is concentrated (Wyner et al., 2017). Therefore, existing frameworks bypass the finite-sample filtering mechanisms required

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

to theoretically calibrate resampling ratio α. To shift from unguided trial-and-error to a theoretically principled adaptation, we must answer two mechanistic questions:

- ***Exactly which functional components of a non-parametric model's variance are reduced by resampling, and which are preserved?***
- ***By what exact algebraic proportion is each interaction component attenuated in finite samples?***

In this paper, we answer these questions by replacing overlap-based covariance calculations with a probabilistic indicator-decoupling argument over individual interaction subsets. The resulting identity shows that subagging is not a uniform variance reducer, but a deterministic spectral filter: it preserves low-order structural signals while attenuating the c-th order interaction variance by an exact factor bounded by α^c. This explains why fixed defaults such as $\alpha \approx 0.632$ can under-regularize high-capacity interpolators, which instead require smaller α to suppress high-order noise. Our main contributions are summarized as follows:

- **Exact Finite-Sample Variance Identity:** We derive the first exact variance decomposition for subagging (Theorem 3.1). This identity is applicable to any symmetric base learner for any finite sample size n and subsample size $k \leq n$, without requiring smoothness assumptions or asymptotic limits.
- **The Spectral Filtering Law:** We analytically characterize the exact attenuation factor $\gamma_c(n, k)$ for each interaction order c (Theorem 3.2). We prove that subagging attenuates the c-th order variance component by a combinatorial factor bounded by, and asymptotically converging to, α^c.
- **Structural Shift of Optimal Regularization:** By integrating the exact variance identity with non-parametric bias rates, we establish a monotone comparative statics result (Theorem 4.5). We show that the optimal subsampling ratio α^* is monotonically non-increasing with the learner's interaction complexity. This formalizes the transition in the bias-variance trade-off: low-capacity smoothers require larger α for bias control, whereas models dominated by high-order interactions necessitate smaller α to suppress noise. Consequently, this result bridges finite-sample resampling heuristics with the asymptotic consistency requirement $\alpha \to 0$.
- **Complexity-Guided Adaptive Subsampling and Empirical Validation:** To operationalize this theory without prohibitive full-spectrum estimations, we propose Complexity-Guided Adaptive Subsampling (CGAS). We empirically validate our structural shift theorem, demonstrating that dynamically calibrating α to the learner's structural capacity systematically improves finite-sample generalization over static baselines.

2 Mathematical Framework

In this section, we develop the rigorous foundation for exact finite-sample analysis. We model the base learner as a symmetric measurable function and represent the subagged ensemble as an infinite-order U-statistic. To isolate variance reduction from structural approximation, we fix a test point $x \in \mathcal{X}$ and study point-wise variance. We provide a comprehensive summary of global symbols and notations in Table 2 in Appendix A.

2.1 Problem Setup and the Subagging Estimator

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be the sample space, and let $\mathcal{D}_n = \{Z_1, \ldots, Z_n\}$ denote n i.i.d. observations drawn from an unknown distribution P on $\mathcal{Z}$. A base learner trained on k samples, where $1 \leq k \leq n$, is represented by a measurable function $h : \mathcal{X} \times \mathcal{Z}^k \to \mathbb{R}$. We impose the following standard regularity assumption on the base learner:

Assumption 2.1 (Symmetry and Square-Integrability (Hoeffding, 1992; Wager, 2014))**.** For any fixed target point $x \in \mathcal{X}$, the base learner $h(x; \cdot)$ satisfies the following regularity conditions: (i) **Permutation Invariance (Symmetry):** The base learner treats the training data as an exchangeable multi-set. Specifically, for any permutation π in the symmetric group $\mathcal{S}_k$ and for all $(z_1, \ldots, z_k) \in$

$\mathcal{Z}^k$ (Serfling, 2009), it holds strictly that

$$h(x; z_1, \ldots, z_k) = h(x; z_{\pi(1)}, \ldots, z_{\pi(k)}).$$

(ii) **Square-Integrability:** The function h is square-integrable with respect to the product measure $P^{\otimes k}$, ensuring a finite second moment (Van der Vaart, 2000):

$$\mathbb{E}_{Z_{1:k} \sim P^{\otimes k}} \left[h(x; Z_1, \ldots, Z_k)^2 \right] < \infty.$$

Assumption 2.1 is standard in U-statistic analyses of ensemble methods (Hoeffding, 1992; Lee, 2019). Condition (i) captures base learners whose predictions depend on the training sample as an unordered set, such as decision trees and nearest-neighbor rules (Mentch & Hooker, 2016; Cover & Hart, 1967; Loh, 2011), while Condition (ii) ensures the finite second moment required for the Hoeffding–ANOVA variance decomposition (Bühlmann & Yu, 2002; Wager & Athey, 2018); it holds trivially for bounded-output classifiers. Randomized or order-dependent learners can be handled by replacing the algorithm with its symmetrized conditional mean, thereby separating structural resampling variance from algorithmic randomness by the law of total variance; the formal construction is given in Appendix B.1.

Assumption 2.1 holds for most unweighted ensemble base learners, such as standard decision trees and nearest neighbors, provided the algorithm does not depend on the specific ordering of the training data (Cover & Hart, 1967; Loh, 2011). We analyze the *Subagging* estimator. For a given subsample size $k \leq n$, the subagged estimator $\hat{F}_{n,k}(x)$ aggregates the predictions of base learners trained on all possible subsets of size k drawn without replacement from $\mathcal{D}_n$ (Mentch & Hooker, 2016; Wager & Athey, 2018):

$$\hat{F}_{n,k}(x) = \binom{n}{k}^{-1} \sum_{S \in \binom{[n]}{k}} h(x; Z_S), \tag{1}$$

where $[n] = \{1, 2, \ldots, n\}$ and $\binom{[n]}{k}$ denotes the set of all k-combinations of $[n]$. Z_S represents the subset of random variables indexed by S. We focus on sampling without replacement because it preserves the subset-wise orthogonality needed for an exact finite-sample decomposition; the distinction from bootstrap sampling is discussed in Appendix B.2.

2.2 The Hoeffding-ANOVA Decomposition and Spectral Components

Eq. (1) reveals that $\hat{F}_{n,k}(x)$ is formulated as a U-statistic of order k with the base learner h as its kernel. To decouple its variance, we invoke the Hoeffding-ANOVA decomposition (Hoeffding, 1992; Efron & Stein, 1981). We project the base learner h into a sum of mutually orthogonal functions of increasing interaction orders.

Definition 2.2 (Canonical ANOVA Projections (Efron & Stein, 1981))**.** Let $\theta(x) = \mathbb{E}_{Z_{1:k} \sim P^{\otimes k}}[h(x; Z_1, \ldots, Z_k)]$ denote the expected prediction of the base learner, representing the 0-th order baseline component. For any $1 \leq c \leq k$, the c-th order canonical projection h_c is recursively defined as:

$$h_c(x; z_1, \ldots, z_c) = \mathbb{E}\big[h(x; Z_{1:k}) \mid Z_1 = z_1, \ldots, Z_c = z_c\big] - \sum_{j=1}^{c-1} \sum_{J \in \binom{[c]}{j}} h_j(x; z_J) - \theta(x), \tag{2}$$

where z_J denotes the specific subset of variables $\{z_i : i \in J\}$, and the conditional expectation integrates out the remaining $k - c$ variables under the product measure $P^{\otimes(k-c)}$.

By this recursive construction, h_c isolates the c-th order pure interaction among the training samples, completely stripped of any lower-order additive main effects. For illustration, the first-order (main effect) and second-order (interaction) kernels are strictly formulated as:

$$\begin{aligned} h_1(x; z_1) &= \mathbb{E}[h \mid Z_1 = z_1] - \theta(x), \\ h_2(x; z_1, z_2) &= \mathbb{E}[h \mid Z_1 = z_1, Z_2 = z_2] - h_1(x; z_1) - h_1(x; z_2) - \theta(x). \end{aligned}$$

As a direct consequence of Definition 2.2, these projections exhibit fundamental orthogonality properties, which we restate below for completeness.

Lemma 2.3 (Strong Degeneracy and Orthogonality (Hoeffding, 1992)). *For any $1 \le c \le k$, the c-th order canonical projection h_c defined in Eq. (2) satisfies the strong degeneracy property: the expectation with respect to any single argument evaluates to zero $P^{\otimes(c-1)}$-almost surely. That is, for any fixed values $\{z_j\}_{j\neq i}$ and any $i \in \{1,\dots,c\}$,*

$$\mathbb{E}_{Z_i\sim P}\big[h_c(x; z_1,\dots,z_{i-1}, Z_i, z_{i+1},\dots,z_c)\big] = 0.$$

Therefore, the canonical projections associated with different index subsets are mutually orthogonal. For any two distinct subsets $C_1, C_2 \subseteq \{1,\dots,k\}$ such that $C_1 \neq C_2$, the covariance between their respective projections is exactly zero:

$$\mathbb{E}_{Z_{1:k}\sim P^{\otimes k}}\big[h_{|C_1|}(x; Z_{C_1}) h_{|C_2|}(x; Z_{C_2})\big] = 0.$$

Lemma 2.3 provides the fundamental mathematical mechanism that allows us to decompose the exact variance of the ensemble into a finite sum of independent variance components without cross-terms. The full proof of Lemma 2.3 is provided in Appendix C. We now quantify the magnitude of these orthogonal components.

Definition 2.4 (ANOVA / Hoeffding Variance Component (Efron & Stein, 1981; Xu et al., 2024)). For $1 \le c \le k$, the c-th order variance component (also known as the *Hoeffding spectrum* or *Sobol index*) of the base learner at target point x is defined as:

$$\zeta_c(x) = \mathrm{Var}\big(h_c(x; Z_1,\dots,Z_c)\big) \equiv \mathbb{E}_{Z_{1:c}\sim P^{\otimes c}}\big[h_c(x; Z_1,\dots,Z_c)^2\big].$$

This quantity measures the variance contributed purely by c-sample interactions; further interpretation is provided in Appendix B.3. Because Lemma 2.3 guarantees the strict mutual orthogonality of all canonical projections, the cross-covariance terms vanish completely. Therefore, the total variance of the base learner evaluates exactly to a weighted sum of its Hoeffding components. We formalize this fundamental property as the following lemma.

Lemma 2.5 (Exact Base Variance Decomposition (Hoeffding, 1992)). *Under Assumption 2.1, the total variance of a single base learner trained on k samples is decoupled into the finite sum:*

$$\mathit{Var}_{Z_{1:k}\sim P^{\otimes k}}\big(h(x; Z_{1:k})\big) = \sum_{c=1}^{k} \binom{k}{c} \zeta_c(x), \tag{3}$$

where the combinatorial weight $\binom{k}{c}$ accounts for the number of distinct c-element subsets drawn from the k training samples.

The full proof of Lemma 2.5 is provided in Appendix D. Eq. (3) represents the functional complexity of the base learner through its Hoeffding spectrum. Additional interpretation of $\zeta_c(x)$, including its relation to functional ANOVA and Sobol-type decompositions, is given in Appendix B.3. Regularized learners, such as shallow trees or additive smoothers, tend to concentrate this spectrum at low orders (Hastie & Tibshirani, 1987; Mentch & Hooker, 2016; Scornet et al., 2015), whereas high-capacity interpolators can place substantial variance on higher orders (Lin & Jeon, 2006; Hooker, 2007).

3 Exact Variance Decomposition and Spectral Filtering

In this section, we derives the finite-sample transformation applied by subagging to the Hoeffding spectrum of the base learner. We first state the exact variance identity, and then show that its attenuation factors form a deterministic low-pass spectral filter.

3.1 Exact Finite-Sample Variance Identity

The variance of the subagged ensemble $\hat{F}_{n,k}(x)$ in Eq. (1) is governed by the covariance between base learners trained on overlapping subsamples. Let $S_1, S_2 \in \binom{[n]}{k}$ be independent uniform subsamples, and let $\rho = |S_1 \cap S_2|$. A direct covariance calculation conditions on ρ, which follows a hypergeometric distribution, and then combines this random overlap with the Hoeffding interaction order c. This produces a nested hypergeometric-binomial convolution: the coefficient of each $\zeta_c(x)$

is spread across all overlap sizes $\rho \geq c$. Reducing this expression to a closed-form attenuation factor for every c, while retaining all high-order components rather than passing to first-order asymptotics, is the central finite-sample combinatorial difficulty. The full overlap-based expansion is given in Appendix E.1.

Our proof avoids this convolution by changing the counting object. Instead of conditioning on the macroscopic overlap ρ, we fix a c-element interaction subset C and count its joint inclusion probability in S_1 and S_2. This microscopic indicator-decoupling argument uses the independence of the two resampling draws and the orthogonality of the Hoeffding projections to isolate the exact coefficient of each spectrum component. It yields the following finite-sample variance identity.

Theorem 3.1 (Exact Variance Decomposition of Subagging)**.** *Under Assumption 2.1, for any finite sample size $n \geq 1$, subsample size $k \leq n$, and an arbitrary test point $x \in \mathcal{X}$, the variance of the subagged estimator $\hat{F}_{n,k}(x)$ in Eq.* (1) *is exactly decoupled into:*

$$Var(\hat{F}_{n,k}(x)) = \sum_{c=1}^{k} \gamma_c(n,k) \binom{k}{c} \zeta_c(x), \tag{4}$$

where $\zeta_c(x)$ is the c-th order Hoeffding variance component, and the attenuation factor $\gamma_c(n,k)$ for the c-th order spectrum is defined as:

$$\gamma_c(n,k) = \frac{\binom{n-k}{k-c}}{\binom{n}{k}} = \prod_{j=0}^{c-1} \frac{k-j}{n-j}. \tag{5}$$

The full proof of Theorem 3.1 is provided in Appendix F.

Theorem 3.1 is an exact algebraic identity, not an asymptotic approximation. Comparing Eq. (4) with the single-learner decomposition in Eq. (3), subagging leaves the same Hoeffding components $\zeta_c(x)$ but rescales the c-th order contribution by $\gamma_c(n,k)$. Thus variance reduction is not uniform across functional orders. A comparison with first-order asymptotic projection analyses is given in Appendix E.2 (Mentch & Hooker, 2016; Wager & Athey, 2018; Xu et al., 2024).

3.2 The Spectral Filtering Law

To uncover the physical intuition behind the attenuation factor $\gamma_c(n,k)$, it is instructive to examine its behavior in the large-sample regime where the subsampling ratio remains constant. Let $\alpha = k/n \in (0,1]$ denote the subsampling ratio. The following theorem characterizes $\gamma_c(n,k)$ as a geometric weight that governs the variance reduction across different interaction orders.

Theorem 3.2 (The Spectral Filtering Law)**.** *For any finite $n \geq k \geq 1$ and any interaction order $1 \leq c \leq k$, the attenuation factor $\gamma_c(n,k)$ defined in Eq.* (5) *satisfies the following properties:*

- *The attenuation is strictly bounded by the c-th power of the subsampling ratio:*

$$\gamma_c(n,k) = \prod_{j=0}^{c-1} \frac{k-j}{n-j} \leq \left(\frac{k}{n}\right)^c = \alpha^c,$$

  where the equality holds if and only if $c = 1$ or $k = n$.

- *For a fixed interaction order c, as $n \to \infty$ with $k/n \to \alpha \in (0,1]$, the attenuation factor converges to the geometric limit:*

$$\lim_{n\to\infty, \frac{k}{n}\to\alpha} \gamma_c(n,k) = \alpha^c.$$

Consequently, the exact ensemble variance in Eq. (4) *is governed by this geometric attenuation across the Hoeffding spectrum.*

The proof follows by viewing $\gamma_c(n,k)$ as the probability that a fixed c-element interaction subset is included in a uniform k-subsample and then applying the chain rule to the successive inclusion probabilities. The finite population dependence makes each subsequent factor no larger than α; this

dependence vanishes for fixed c in the proportional asymptotic limit. The full proof is provided in Appendix G.

Theorem 3.2 shows that subagging acts as a deterministic low-pass filter on the Hoeffding spectrum. Low-order components are only mildly attenuated, while high-order interactions receive an approximately geometric penalty α^c. Hence high-capacity interpolators, whose variance can be concentrated in large c, require smaller α to suppress high-order noise. This spectral view motivates the bias–variance analysis of the optimal subsampling ratio in Section 4.

4 Finite-Sample Bias-Variance Trade-off and Optimal Regularization

In this section, we combine the exact variance identity with a standard nonparametric bias model to study the finite-sample bias–variance trade-off of subagging. While the spectral filtering law characterizes variance reduction for a fixed subsample size k, choosing k, or equivalently $\alpha = k/n$, also changes the bias of the base learner. We use this trade-off to characterize how the optimal subsampling ratio depends on the learner's Hoeffding interaction spectrum.

4.1 Finite-Sample MSE Bound

Let $f^*(x) = \mathbb{E}[Y \mid X = x]$ and $\theta_k(x) = \mathbb{E}[h(x; Z_{1:k})]$. By linearity of expectation,

$$\mathrm{Bias}(\hat{F}_{n,k}(x)) = \mathbb{E}[\hat{F}_{n,k}(x)] - f^*(x) = \theta_k(x) - f^*(x) \triangleq \mathcal{B}(k). \tag{6}$$

Thus Eq. (6) confirms that subagging does not change the bias of a base learner trained on k samples; it only averages its variance.

Assumption 4.1 (Monotone bias decay)**.** For the fixed point x, the squared bias satisfies $\mathcal{B}(k)^2 \leq B_0 k^{-2\beta}$, where $B_0 > 0$ is a constant depending on the intrinsic complexity of the target function, and $\beta > 0$ governs the non-parametric learning rate.

Polynomial bias envelopes of this form are standard in nonparametric regression: for Hölder-smooth functions, minimax theory gives algebraic rates depending on smoothness and dimension (Stone, 1982; Györfi et al., 2002). Related algebraic bias rates also appear in analyses of random forests and tree-based learners (Biau, 2012; Wager & Athey, 2018); see Appendix H.1 for additional discussion.

Lemma 4.2 (Finite-Sample MSE Bound)**.** *Under Assumption 4.1, for any $n \geq k \geq 1$ and fixed $x \in \mathcal{X}$,*

$$\mathit{MSE}(\hat{F}_{n,k}(x)) \leq B_0 k^{-2\beta} + \sum_{c=1}^{k} \gamma_c(n,k) \binom{k}{c} \zeta_c(x). \tag{7}$$

The proof is the standard bias–variance decomposition (Geman et al., 1992; Hastie et al., 2009) with Theorem 3.1 substituted for the variance term; the full proof is given in Appendix I. Eq. (7) displays the trade-off: increasing k reduces the bias term but weakens the spectral attenuation of high-order variance.

4.2 The Structural Dependence of the Optimal Subsampling Path

To optimize the MSE bound in Eq. (7), we move from the discrete subsample size k to the continuous ratio $\alpha = k/n$. This requires separating the learner's intrinsic interaction complexity from the variable subsample size.

Assumption 4.3 (Structural Interaction Capacity)**.** For a given learning algorithm and evaluation point x, there exists $M \geq 1$ such that $\zeta_c(x) = 0$ for all $c > M$. Low-complexity smoothers have small effective M, whereas high-capacity interpolators can have substantial spectral mass at large interaction orders.

Assumption 4.3 lets us define a continuous optimization objective over α without a variable summation limit. The following lemma is the bridge from the discrete finite-sample MSE bound to comparative statics.

Lemma 4.4 (Continuous Surrogate MSE Envelope)**.** *Suppose Assumptions 4.1 and 4.3 hold. For* $\alpha \in [M/n, 1]$*, define*

$$\mathcal{M}(\alpha; \boldsymbol{\zeta}, M) \triangleq B_0(\alpha n)^{-2\beta} + \sum_{c=1}^{M} \alpha^c \frac{\Gamma(\alpha n + 1)}{\Gamma(c+1)\Gamma(\alpha n - c + 1)} \zeta_c(x).$$

Then $\mathcal{M}(\alpha; \boldsymbol{\zeta}, M)$ *is a differentiable upper envelope of the finite-sample MSE bound in Eq.* (7)*.*

The proof uses $\gamma_c(n, \alpha n) \leq \alpha^c$ from Theorem 3.2 and the Gamma-function extension of the binomial coefficient; the constraint $\alpha n \geq M$ keeps the Gamma arguments away from singularities. The full proof is given in Appendix J. The stationary condition for this surrogate is nonlinear and generally has no useful closed form; see Appendix H.2. We therefore characterize the optimizer through monotone comparative statics.

Theorem 4.5 (Structural Shift of Optimal Regularization)**.** *Suppose two base learning algorithms,* $h^{(1)}$ *and* $h^{(2)}$*, have identical bias parameters* (B_0, β) *and Hoeffding spectra* $\boldsymbol{\zeta}^{(1)} = \{\zeta_c^{(1)}\}_{c=1}^M$ *and* $\boldsymbol{\zeta}^{(2)} = \{\zeta_c^{(2)}\}_{c=1}^M$*. Assume there exists a critical order* c^* *such that* $\zeta_c^{(2)} = \zeta_c^{(1)}$ *for* $c < c^*$*,* $\zeta_c^{(2)} \geq \zeta_c^{(1)}$ *for* $c \geq c^*$*, with strict inequality for at least one* $c \geq c^*$*. Then the optimal subsampling ratios satisfy* $\alpha^*_{(2)} \leq \alpha^*_{(1)}$*.*

The proof applies Topkis's monotone comparative statics theorem (Topkis, 1978; 1998). The key observation is that shifting variance mass toward higher orders makes the variance part of $\mathcal{M}(\alpha; \boldsymbol{\zeta}, M)$ more costly at larger α, so the negative objective has decreasing differences in $(\alpha, \boldsymbol{\zeta})$. Thus the optimal ratio is non-increasing as the learner's spectrum becomes more high-order. The full proof is given in Appendix K.

Theorem 4.5 formalizes the structural limitation of static resampling rules such as $\alpha \approx 0.632$, which arise from bootstrap inclusion probabilities rather than the finite-sample bias–variance tradeoff (Breiman, 1996b; Efron & Tibshirani, 1997; Bühlmann & Yu, 2002). Low-complexity learners favor larger α to control bias, whereas high-capacity interpolators benefit from smaller α because the factor α^c suppresses high-order variance. A fuller discussion of these regimes and the strict interior-solution corollary is provided in Appendix H.3.

4.3 Illustration: A Bi-Modal Spectrum

To make the preceding comparative statics concrete, consider an illustrative learner whose variance is concentrated at two interaction orders: a low-order signal component at $c = 1$ and a high-order noise component at $c = M$. The spectrum is normalized so that the standalone base learner has total variance $\sigma_1^2 + \sigma_M^2$. Substituting this spectrum into the filtered MSE bound gives

$$\begin{aligned} \mathcal{M}(\alpha; M) &= B_0(\alpha n)^{-2\beta} + \gamma_1(\alpha n, n)\sigma_1^2 + \gamma_M(\alpha n, n)\sigma_M^2 \\ &= B_0 n^{-2\beta}\alpha^{-2\beta} + \sigma_1^2\alpha + \sigma_M^2\alpha^M\left(1 + \mathcal{O}\left(\frac{M^2}{n}\right)\right). \end{aligned} \tag{8}$$

Eq. (8) distills the core mechanics of ensemble learning: the low-order signal variance is attenuated linearly, while high-order noise receives the much sharper geometric penalty α^M. Balancing the dominant high-order variance term with the bias derivative yields $\alpha^* \asymp (\sigma_M^2)^{-1/(M+2\beta)}$ in the high-complexity regime. Hence larger high-order noise pushes the optimal subsampling ratio downward, matching Theorem 4.5. The full derivation and regime discussion are given in Appendix L.

5 Algorithm Design and Empirical Verification

In this section, we introduce Complexity-Guided Adaptive Subsampling (CGAS) to test the predicted structural shift of the optimal subsampling ratio α^*. CGAS uses a structural capacity proxy of the base learner to restrict the search space for α, following the spectral filtering mechanism in Theorem 3.2 and the comparative statics in Theorem 4.5. We evaluate CGAS across low-complexity smoothers and high-complexity interpolators, comparing the selected α^* and generalization performance against fixed-rate subagging baselines.

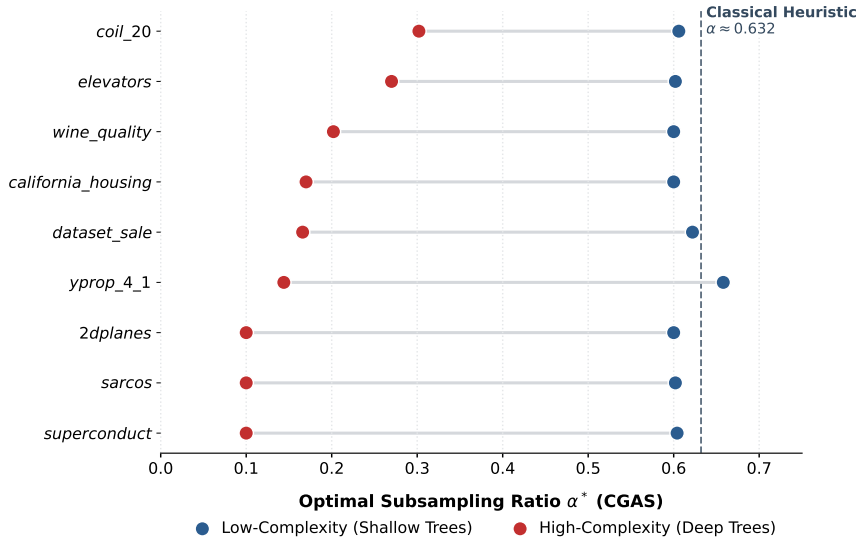

Figure 1: Selected subsampling ratios α^* across nine datasets.

5.1 CGAS DESIGN.

Theorem 4.5 suggests that learners with heavier high-order spectra should use smaller subsampling ratios. Since estimating the full Hoeffding spectrum $\{\zeta_c\}_{c=1}^{k}$ is computationally prohibitive (Wager & Athey, 2018; Xu et al., 2024), CGAS uses a structural complexity proxy, such as maximum tree depth $d_{\max}$, to restrict the candidate grid for α and then selects α^* by internal cross-validation. For low-complexity learners, CGAS searches larger subsampling ratios because aggressive subsampling mainly increases bias. For high-complexity learners, it searches smaller ratios to impose stronger geometric attenuation on high-order variance. In our implementation, the low-complexity grid is $\{0.6, 0.7, 0.8, 0.9, 0.95\}$ and the high-complexity grid is $\{0.1, 0.2, 0.3, 0.4\}$. The selected ratio is then used to train the final subagged ensemble. The detailed procedure and pseudocode are given in Appendix M.1.

5.2 EXPERIMENTAL PROTOCOL.

We evaluate CGAS on nine OpenML regression datasets spanning different sample sizes and feature dimensions. To test the predicted structural shift, we use two controlled regimes: shallow decision trees with $d_{\max} = 3$ on the original targets as low-complexity smoothers, and unpruned decision trees with additive Gaussian label noise scaled to $1.5\sigma_y$ as high-complexity interpolators. CGAS is compared with fixed-rate subagging baselines using $\alpha = 0.632$ and $\alpha = 0.8$, with `n_estimators=100` for all methods. We report out-of-sample MSE and MAE under repeated cross-validation, and assess significance using a one-sided Wilcoxon signed-rank test. Full implementation details, evaluation protocol, and environment information are provided in Appendix M.2.

5.3 EMPIRICAL RESULTS AND ANALYSIS

The empirical analysis focuses on two main-text questions: whether the selected subsampling ratio α^* shifts with learner complexity (Section 5.3.1), and whether CGAS improves over fixed-rate subagging baselines (Section 5.3.2). Additional robustness results for alternative symmetric learners are reported in Appendix N.1, and the extension to standard bootstrap random forests is reported in Appendix N.2.

5.3.1 STRUCTURAL SHIFT OF THE OPTIMAL SUBSAMPLING RATIO

Figure 1 shows that the subsampling ratio selected by CGAS shifts systematically with learner complexity. For shallow trees, α^* remains near the classical range, roughly 0.60–0.65. For unpruned

noisy trees, α^* moves to much smaller values, typically 0.1–0.3. This pattern is consistent with Theorem 4.5: low-complexity learners need larger subsamples to control bias, whereas high-complexity interpolators benefit from stronger attenuation of high-order variance.

Table 1: Out-of-sample MSE and MAE across various datasets under two distinct functional complexity regimes. The best performance metrics in each regime are highlighted in **bold**.

Datasets	Metric	Regime 1 (Low-Complexity)			Regime 2 (High-Complexity)		
		0.632	0.8	CGAS	0.632	0.8	CGAS
yprop_4_1	MSE	0.9672*	0.9686*	**0.9670**	1.0192*	1.0427*	**0.9923**
	MAE	**0.6605**	0.6610*	**0.6605**	0.7851*	0.7936*	**0.7745**
wine_quality	MSE	0.7231*	0.7276*	**0.7221**	0.9326*	0.9821*	**0.8843**
	MAE	0.6671*	0.6695*	**0.6666**	0.7663*	0.7859*	**0.7457**
superconduct	MSE	0.2881*	0.2908*	**0.2876**	0.8628*	0.9650*	**0.7468**
	MAE	0.3745*	0.3763*	**0.3742**	0.7398*	0.7603*	**0.6886**
dataset_sale	MSE	0.4051*	0.4077*	**0.4043**	0.8538*	0.8973*	**0.8019**
	MAE	**0.3825**	0.3832*	**0.3825**	0.7350*	0.7530*	**0.7106**
coil_20	MSE	0.3602*	0.3762*	**0.3582**	0.7586*	0.7826*	**0.7314**
	MAE	0.4546*	0.4586*	**0.4542**	0.6956*	0.7069*	**0.6839**
elevators	MSE	0.4809*	0.4896*	**0.4796**	0.7902*	0.8078*	**0.7726**
	MAE	0.4575*	0.4614*	**0.4570**	0.7080*	0.7160*	**0.7002**
california_housing	MSE	0.4488*	0.4569*	**0.4472**	0.7979*	0.8125*	**0.7832**
	MAE	0.5021*	0.5068*	**0.5012**	0.7107*	0.7172*	**0.7048**
sarcos	MSE	0.3741*	0.3804*	**0.3729**	0.7796*	0.8079*	**0.7229**
	MAE	0.4573*	0.4619*	**0.4563**	0.7036*	0.7157*	**0.6779**
2dplanes	MSE	0.1793*	0.1831*	**0.1786**	0.9314*	1.0197*	**0.7387**
	MAE	0.3446*	0.3484*	**0.3438**	0.7681*	0.8028*	**0.6856**

Note: All baselines marked with $*$ are significantly different from CGAS according to the Wilcoxon signed-rank test, where * denotes $p < 0.001$.

5.3.2 Performance Evaluation Against Fixed-Rate Subagging

Table 1 compares CGAS with fixed-rate subagging using $\alpha = 0.632$ and $\alpha = 0.8$, providing consistent empirical support for the theoretical mechanisms in Theorem 3.2 and Theorem 4.5. In the high-complexity regime, fixed ratios under-regularize unpruned noisy trees, whereas CGAS selects smaller α and achieves lower MSE and MAE across all datasets; for example, it reduces MSE by 20.7% relative to $\alpha = 0.632$ on *2dplanes*. In the low-complexity regime, CGAS selects larger ratios and avoids the bias inflation caused by overly aggressive subsampling, yielding comparable or slightly better performance than the fixed baselines. These results support the view that α should be treated as a complexity-dependent regularization parameter, rather than a universal constant.

6 Conclusion

This work derived the first exact finite-sample variance identity for subagging and showed that its resampling operator acts as a deterministic spectral filter over the Hoeffding spectrum. The resulting attenuation factor scales approximately as α^c, explaining why high-capacity interpolators with high-order variance often require smaller subsampling ratios than static defaults such as $\alpha \approx 0.632$. Building on this insight, CGAS adaptively calibrates α using learner complexity and improves over fixed-rate baselines. The exact identity applies to sampling without replacement; extending comparable finite-sample decompositions to bootstrap sampling remains an important direction for future work.

AI USE STATEMENT

In this work, we did not use generative AI tools for tasks requiring disclosure such as developing the theoretical framework, formulating mathematical claims, proving theorems, designing experiments, implementing methods, processing data, or interpreting empirical results. We used generative AI tools only for manuscript writing support, including language polishing, improving readability, condensing text, and reorganizing presentation. All AI-assisted text was reviewed and edited by the authors. The authors verified the mathematical statements, proofs, experimental results, and all scientific claims independently, and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work is theoretical and methodological, and does not involve human subjects, private or sensitive data, dataset release, surveillance, security-sensitive applications, or deployment in high-stakes domains. We do not identify specific ethical concerns beyond standard requirements of research integrity, reproducibility, and accurate reporting.

REPRODUCIBILITY STATEMENT

This work is primarily theoretical. We provide the assumptions, definitions, and complete proofs of all theoretical results in the appendix. For the empirical part, the appendix reports the CGAS pseudocode, dataset information, complexity regimes, hyperparameter grids, evaluation protocol, environmental setting, and implementation details. The CGAS procedure is simple and fully specified by the pseudocode, making the experiments directly reproducible from the reported configuration.

REFERENCES

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

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

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996a.

Leo Breiman. Out-of-bag estimation: Technical report. *Department of Statistics, University of California*, 1996b.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001a.

Leo Breiman. Using iterated bagging to debias regressions. *Machine learning*, 45(3):261–277, 2001b.

Peter Bühlmann and Bin Yu. Analyzing bagging. *The annals of Statistics*, 30(4):927–961, 2002.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. Why do random forests work? understanding tree ensembles as self-regularizing adaptive smoothers. *arXiv preprint arXiv:2402.01502*, 2024.

Aaron S Edlin and Chris Shannon. Strict monotonicity in comparative statics. *Journal of Economic Theory*, 81(1):201–219, 1998.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pp. 586–596, 1981.

Bradley Efron and Robert Tibshirani. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334. Springer, 1992.

Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of computational and graphical statistics*, 16(3):709–732, 2007.

Alan J Lee. *U-statistics: Theory and Practice*. Routledge, 2019.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

Gilles Louppe. *Understanding random forests: From theory to practice*. Universite de Liege (Belgium), 2014.

Gonzalo Martínez-Muñoz and Alberto Suárez. Out-of-bag estimation of the optimal sample size in bagging. *Pattern recognition*, 43(1):143–152, 2010.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016.

Nathaniel S O'Connell. Random forests as statistical procedures: Design, variance, and dependence. *arXiv preprint arXiv:2602.13104*, 2026.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 8(4):e1249, 2018.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Richard P Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in advanced mathematics*, 2011.

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

Donald M Topkis. Minimizing a submodular function on a lattice. *Operations research*, 26(2): 305–321, 1978.

Donald M Topkis. *Supermodularity and complementarity*. Princeton university press, 1998.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Stefan Wager. Asymptotic theory for random forests. *arXiv preprint arXiv:1405.0352*, 2014.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48):1–33, 2017.

Tianning Xu, Ruoqing Zhu, and Xiaofeng Shao. On variance estimation of random forests with infinite-order u-statistics. *Electronic Journal of Statistics*, 18(1):2135–2207, 2024.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhengze Zhou, Lucas Mentch, and Giles Hooker. V-statistics and variance estimation. *Journal of Machine Learning Research*, 22(287):1–48, 2021.

Zhi-Hua Zhou. Ensemble learning. In *Machine learning*, pp. 181–210. Springer, 2021.

A NOTATION

To facilitate a rigorous finite-sample analysis and ensure mathematical clarity, we formalize the primary notation used throughout this paper in Table 2.

Table 2: Summary of Global Symbols and Notations

Symbol	Definition	Domain/Space
1. Sets and Spaces		
$\mathcal{X}, \mathcal{Y}$	Feature input space and target continuous output space	-
$\mathcal{Z}$	The measurable product space $\mathcal{X} \times \mathcal{Y}$	-
$\mathcal{D}_n$	The finite training dataset $\{Z_1, \dots, Z_n\}$	$\mathcal{Z}^n$
$[n], [k]$	The integer index sets $\{1, 2, \dots, n\}$ and $\{1, 2, \dots, k\}$	$\mathbb{Z}^+$
$\binom{[n]}{k}, \binom{[n]}{c}$	The collection of all k-element or c-element subsets drawn from $[n]$	$2^{[n]}$
Ω	The discrete sample space of all possible subsamples of size k	$2^{[n]}$
$\mathcal{S}_k$	The symmetric group containing all permutations of k elements	-
$\mathcal{P}(C)$	The Boolean lattice (power set) of interaction index subset C	2^C
$\mathcal{A}_0, \mathcal{A}_1$	Sub-lattice partitions of $\mathcal{P}(C)$ based strictly on the exclusion ($\mathcal{A}_0$) or inclusion ($\mathcal{A}_1$) of a target integration index i	$\subset 2^C$
$\mathcal{H}$	The functional space of all valid Hoeffding variance spectra	$\mathbb{R}^k_{\geq 0}$
2. Data and Random Variables		
Z_i, z_i	Independent and identically distributed (i.i.d.) random variable and its realization	$\mathcal{Z}$
S, S_1, S_2	Index subsets representing distinct subsamples of cardinality k	$\binom{[n]}{k}$
C, J	Arbitrary interaction index subsets, typically of cardinality c and j	$\binom{[n]}{c}, 2^C$
C_1, C_2	Distinct interaction index subsets used to prove the mutual orthogonality of Hoeffding projections ($C_1 \neq C_2$)	$2^{[k]}$
A, B	Arbitrary non-empty index subsets of the training sample used in the exact algebraic variance expansion of the base learner	$2^{[k]} \setminus \{\emptyset\}$

Continued on next page

Table 2 – continued from previous page

Symbol	Definition	Domain/Space
Z_S, z_J	The specific subset of training random variables indexed by S or J	$\mathcal{Z}^{\lvert S\rvert}, \mathcal{Z}^{\lvert J\rvert}$
ξ	Internal algorithmic randomness (e.g., random feature splits)	-
π	Random permutation of training data indices drawn uniformly	$\mathcal{S}_k$
E_j	The event that the j-th element of interaction target C is included in subsample S	-
3. Scalars and Indices		
n, k	Total sample size and the subsample size drawn without replacement	$\mathbb{Z}^+$
c, m	Strict order of multi-sample interaction ($1 \leq c \leq k$)	$\mathbb{Z}^+$
ρ	The cardinality of intersection between two subsamples ($\rho = \lvert S_1 \cap S_2\rvert$)	$\{0, \ldots, k\}$
M	The maximum finite interaction depth (cutoff order) / Structural capacity proxy	$\mathbb{Z}^+$
α	The continuous subsampling ratio, defined as $\alpha = k/n$	$(0, 1]$
α^*	The optimal subsampling regularization parameter	$(0, 1]$
B_0, β	Strict upper-bound constants for the non-parametric bias decay rate	$\mathbb{R}^+$
d	The dimensionality of the input feature space $\mathcal{X}$	$\mathbb{Z}^+$
e	The smoothness exponent of a Hölder-continuous target function	$\mathbb{R}^+$
p	Statistical p-value utilized in empirical hypothesis testing (e.g., Wilcoxon signed-rank test)	$(0, 1)$
Q	Algebraic bias decay exponent specific to random trees, dependent on the number of active features	$\mathbb{R}^+$
σ_1^2, σ_M^2	Invariant total variance contributions of the signal and high-order overfitting noise in the bi-modal case study	$\mathbb{R}^+ \cup \{0\}$
$d_{\max}$	Maximum tree depth used as an algorithmic proxy for structural complexity	$\mathbb{Z}^+ \cup \{\infty\}$
K	Number of internal cross-validation folds in CGAS algorithm	$\mathbb{Z}^+$
$B_{\text{search}}, B_{\text{final}}$	The number of estimators in the pilot and final subagged ensembles	$\mathbb{Z}^+$
4. Functions, Operators, and Estimators		
$h(x; Z_S)$	Deterministic, permutation-invariant base learning algorithm trained on Z_S	$\mathbb{R}$
$\mathcal{A}(x; Z_{1:k}, \xi, \pi)$	Randomized or order-dependent base learning algorithm prior to marginalization	$\mathbb{R}$
$\hat{F}_{n,k}(x)$	Subagged ensemble estimator (incomplete, infinite-order U-statistic)	$\mathbb{R}$
$f^*(x)$	The true target function representing the conditional expectation $\mathbb{E}[Y \lvert X = x]$	$\mathbb{R}$
$\theta(x), \theta_k(x)$	Marginalized expected prediction of a base learner trained on k samples	$\mathbb{R}$
$h_c(x; z_C)$	The c-th order canonical Hoeffding projection kernel (pure interactions)	$\mathbb{R}$
$E_J(z_J)$	Marginalized conditional expectation integrating out remaining variables	$\mathbb{R}$
$H_c(\alpha)$	The variance inflation multiplier for the c-th order spectrum	$\mathbb{R}^+$
$f(\alpha, \boldsymbol{\zeta})$	The continuous objective function $f(\alpha, \boldsymbol{\zeta}) = -\mathcal{M}(\alpha; \boldsymbol{\zeta})$ used for submodularity in Topkis's Theorem	$\mathbb{R}$
$\Gamma(z)$	The Gamma function utilized for the analytic continuation of discrete factorials	$\mathbb{R}$

Continued on next page

Table 2 – continued from previous page

Symbol	Definition	Domain/Space
$\mathcal{O}(\cdot)$	Big-O notation, representing the asymptotic upper bound of a functional growth rate	-
5. Measures, Variance Components, and Objectives		
$P, P^{\otimes k}, P^n$	The unknown data-generating probability measure and its product measures	-
$\mathcal{U}, \mathbb{P}_{\mathcal{U}}$	The uniform probability measure over the discrete subsampling space Ω	-
$\mathbb{E}$, Var, Cov	Mathematical expectation, variance, and covariance operators	-
$\mathbb{I}_{\{\cdot\}}$	The indicator function isolating structural inclusion events	$\{0, 1\}$
$\zeta_c(x)$	The c-th order Hoeffding variance component (Hoeffding spectrum / Sobol index)	$\mathbb{R}^+ \cup \{0\}$
$\boldsymbol{\zeta}$	The partially ordered variance complexity spectrum sequence $\{\zeta_c\}_{c=1}^{M}$	$\mathbb{R}^M$
$\gamma_c(n, k)$	The exact combinatorial attenuation factor (spectral filter transfer function)	$(0, 1]$
$\mathcal{B}(k)$	The inherent finite-sample bias, strictly equal to $\theta_k(x) - f^*(x)$	$\mathbb{R}$
$\mathcal{M}(\alpha; \boldsymbol{\zeta}, M)$	The continuous finite-sample Mean Squared Error (MSE) upper-bounding envelope	$\mathbb{R}^+$
$\succeq$	The partial order defined on the functional space of spectra $\mathcal{H}$	-

B Additional Details for Section 2

B.1 Randomized or Order-Dependent Learners

Assumption 2.1 is stated for deterministic permutation-invariant learners. For a randomized or order-dependent algorithm $\mathcal{A}(x; Z_{1:k}, \xi, \pi)$, let $\xi \sim P_\xi$ denote internal algorithmic randomness, such as random feature choices, and let $\pi \sim \mathrm{Unif}(\mathcal{S}_k)$ denote an independent random permutation of the input order. Define the marginalized structural predictor

$$h(x; Z_{1:k}) = \mathbb{E}_{\xi,\pi}\left[\mathcal{A}(x; Z_{1:k}, \xi, \pi) \mid Z_{1:k}\right].$$

This function is permutation invariant by construction. By the law of total variance, the prediction variance can be decomposed into the variance of the marginalized structural component and the conditional variance due to algorithmic randomness (Wager & Athey, 2018). The finite-sample identities in the main text analyze the former, isolating the effect of combinatorial resampling.

B.2 Bootstrap versus Subagging

The exact decomposition in this paper is for sampling without replacement. Bootstrap sampling with replacement allows repeated observations within the same resample, creating diagonal collisions in the kernel arguments (Efron & Tibshirani, 1994). These repeated indices couple different interaction orders and prevent the direct subset-wise orthogonality used by the Hoeffding–ANOVA decomposition. When $k \ll n$, bagging and subagging are asymptotically close in several regimes (Bühlmann & Yu, 2002; Wager, 2014), but the exact finite-sample identity derived here relies on the without-replacement structure.

B.3 Interpretation of the Hoeffding Spectrum

The coefficient $\zeta_c(x)$ measures the variance contribution of the pure c-way interaction among training samples after all lower-order effects have been removed. This is closely related to functional ANOVA and Sobol-type decompositions (Efron & Stein, 1981; Hooker, 2007; Xu et al., 2024). A learner whose prediction depends mainly on additive or low-order structure has most of its variance concentrated at small c. Examples include shallow trees, additive smoothers, or other strongly regularized procedures (Hastie & Tibshirani, 1987; Mentch & Hooker, 2016; Scornet et al., 2015).

By contrast, highly adaptive interpolators, such as fully grown decision trees, can respond to localized configurations involving many training points, producing non-negligible high-order variance components (Lin & Jeon, 2006). This spectrum is the object transformed by the subagging filter analyzed in Section 3.

C Proof of Lemma 2.3

Proof. Let $C \subseteq \{1, \ldots, k\} = [k]$ be an index subset with $|C| = c \geq 1$. For any $J \subseteq C$, let $E_J(z_J) \triangleq \mathbb{E}[h(x; Z_{1:k}) \mid Z_J = z_J]$ denote the marginalized conditional expectation. By the recursive construction in Eq. (2), the kernel h_c satisfies the relation $E_C(z_C) = \sum_{J \subseteq C} h_{|J|}(x; z_J)$. Applying Möbius inversion over the Boolean lattice of subsets $(\mathcal{P}(C), \subseteq)$ (Stanley, 2011), we obtain the explicit form:

$$h_c(x; z_C) = \sum_{J \subseteq C} (-1)^{|C|-|J|} E_J(z_J). \tag{9}$$

To prove strong degeneracy, fix an arbitrary index $i \in C$ and consider the expectation of h_c over $Z_i \sim P$. Applying $\mathbb{E}_{Z_i}$ to Eq. (9) and invoking the linearity of expectation:

$$\mathbb{E}_{Z_i}[h_c(x; z_C)] = \sum_{J \subseteq C} (-1)^{|C|-|J|} \mathbb{E}_{Z_i}\left[E_J(z_J)\right]. \tag{10}$$

Define the sub-lattice partition $\mathcal{P}(C) = \mathcal{A}_0 \cup \mathcal{A}_1$, where $\mathcal{A}_0 = \{J \in \mathcal{P}(C) : i \notin J\}$ and $\mathcal{A}_1 = \{J \in \mathcal{P}(C) : i \in J\}$. There exists a bijection $\phi : \mathcal{A}_0 \to \mathcal{A}_1$ defined by $\phi(J) = J \cup \{i\}$. We evaluate $\mathbb{E}_{Z_i}[E_J]$ for each case:

1. For $J \in \mathcal{A}_0$: Since $Z_i \notin Z_J$, $E_J(z_J)$ is constant w.r.t. Z_i, thus $\mathbb{E}_{Z_i}[E_J(z_J)] = E_J(z_J)$.
2. For $J \in \mathcal{A}_1$: Let $J = J' \cup \{i\}$ where $J' \in \mathcal{A}_0$. By the tower property of conditional expectation:

$$\mathbb{E}_{Z_i}[E_{J' \cup \{i\}}(z_{J' \cup \{i\}})] = \mathbb{E}_{Z_i}[\mathbb{E}[h \mid Z_{J'}, Z_i]] = \mathbb{E}[h \mid Z_{J'}] = E_{J'}(z_{J'}).$$

Substituting these results and the bijection ϕ into Eq. (10) yields:

$$\begin{aligned}
\mathbb{E}_{Z_i}[h_c(x; z_C)] &= \sum_{J \in \mathcal{A}_0} (-1)^{|C|-|J|} \mathbb{E}_{Z_i}[E_J] + \sum_{J \in \mathcal{A}_1} (-1)^{|C|-|J|} \mathbb{E}_{Z_i}[E_J] \\
&= \sum_{J \in \mathcal{A}_0} (-1)^{|C|-|J|} E_J(z_J) + \sum_{J \in \mathcal{A}_0} (-1)^{|C|-|J \cup \{i\}|} E_J(z_J) \\
&= \sum_{J \in \mathcal{A}_0} \left((-1)^{|C|-|J|} + (-1)^{|C|-|J|-1}\right) E_J(z_J) \\
&= \sum_{J \in \mathcal{A}_0} \left((-1)^{|C|-|J|} - (-1)^{|C|-|J|}\right) E_J(z_J) = 0.
\end{aligned}$$

This confirms that h_c is strongly degenerate P-almost surely for all $i \in C$.

Consider two distinct index subsets $C_1, C_2 \subseteq \{1, \ldots, k\}$ such that $C_1 \neq C_2$. Without loss of generality, assume there exists $i \in C_1 \setminus C_2$. Let $U = (C_1 \cup C_2) \setminus \{i\}$. The covariance evaluates to:

$$\begin{aligned}
\mathbb{E}_{Z_{C_1 \cup C_2}}\left[h_{|C_1|}(x; Z_{C_1}) h_{|C_2|}(x; Z_{C_2})\right] &= \mathbb{E}_{Z_U}\left[\mathbb{E}_{Z_i}\left[h_{|C_1|}(x; Z_{C_1}) h_{|C_2|}(x; Z_{C_2}) \mid Z_U\right]\right] \\
&\overset{(a)}{=} \mathbb{E}_{Z_U}\left[h_{|C_2|}(x; Z_{C_2}) \cdot \mathbb{E}_{Z_i}\left[h_{|C_1|}(x; Z_{C_1}) \mid Z_U\right]\right] \\
&\overset{(b)}{=} \mathbb{E}_{Z_U}\left[h_{|C_2|}(x; Z_{C_2}) \cdot 0\right] = 0,
\end{aligned}$$

where (a) follows since $i \notin C_2$, making $h_{|C_2|}$ independent of Z_i, and (b) follows from the strong degeneracy of $h_{|C_1|}$ w.r.t. Z_i. This completes the proof. □

D PROOF OF LEMMA 2.5

Proof. By reversing the recursive definition of the canonical projections in Eq. (2), the centralized base learner can be expressed as an exact sum over all non-empty subsets of the index set $[k]$:

$$h(x; Z_{1:k}) - \theta(x) = \sum_{\substack{A \subseteq [k] \\ A \neq \emptyset}} h_{|A|}(x; Z_A).$$

By definition, the variance of the base learner is the expectation of its squared centered prediction. Expanding the square of the sum yields:

$$\begin{aligned} \mathrm{Var}\big(h(x; Z_{1:k})\big) &= \mathbb{E}_{Z_{1:k} \sim P^{\otimes k}} \left[\left(\sum_{\substack{A \subseteq [k] \\ A \neq \emptyset}} h_{|A|}(x; Z_A) \right)^2 \right] \\ &= \mathbb{E}_{Z_{1:k} \sim P^{\otimes k}} \left[\sum_{\substack{A \subseteq [k] \\ A \neq \emptyset}} \sum_{\substack{B \subseteq [k] \\ B \neq \emptyset}} h_{|A|}(x; Z_A) h_{|B|}(x; Z_B) \right]. \end{aligned}$$

By the linearity of expectation over finite sums, we interchange the expectation and the double summation:

$$\mathrm{Var}\big(h(x; Z_{1:k})\big) = \sum_{\substack{A \subseteq [k] \\ A \neq \emptyset}} \sum_{\substack{B \subseteq [k] \\ B \neq \emptyset}} \mathbb{E}_{Z_{1:k} \sim P^{\otimes k}} \Big[h_{|A|}(x; Z_A) h_{|B|}(x; Z_B) \Big].$$

According to Lemma 2.3, the canonical projections are mutually orthogonal. For any two distinct subsets $A \neq B$, their cross-covariance evaluates exactly to zero. Consequently, the double summation collapses strictly to its diagonal terms (where $A = B$):

$$\begin{aligned} \mathrm{Var}\big(h(x; Z_{1:k})\big) &= \sum_{\substack{A \subseteq [k] \\ A \neq \emptyset}} \mathbb{E}_{Z_{1:k} \sim P^{\otimes k}} \Big[h_{|A|}(x; Z_A)^2 \Big] \\ &= \sum_{\substack{A \subseteq [k] \\ A \neq \emptyset}} \mathrm{Var}\big(h_{|A|}(x; Z_A)\big). \end{aligned}$$

To evaluate this single summation analytically, we partition the terms based on the subset cardinality $c = |A|$:

$$\mathrm{Var}\big(h(x; Z_{1:k})\big) = \sum_{c=1}^{k} \sum_{\substack{A \subseteq [k] \\ |A| = c}} \mathrm{Var}\big(h_c(x; Z_A)\big).$$

Under Assumption 2.1 and the i.i.d. measure $P^{\otimes k}$, the variance of a canonical projection depends strictly on its interaction order c, independent of the specific indices in subset A. Thus, we substitute $\mathrm{Var}(h_c(x; Z_A)) = \zeta_c(x)$:

$$\mathrm{Var}\big(h(x; Z_{1:k})\big) = \sum_{c=1}^{k} \sum_{\substack{A \subseteq [k] \\ |A| = c}} \zeta_c(x).$$

Since the inner summation iterates over a constant $\zeta_c(x)$ for all possible c-element subsets drawn from $[k]$, it resolves to a simple count. There are exactly $\binom{k}{c}$ such distinct subsets, yielding:

$$\mathrm{Var}\big(h(x; Z_{1:k})\big) = \sum_{c=1}^{k} \binom{k}{c} \zeta_c(x).$$

This completes the proof. □

E Additional Discussion for Theorem 3.1

E.1 Overlap-Based Hypergeometric Form

An equivalent but less transparent derivation conditions on the overlap $\rho = |S_1 \cap S_2|$ of two independent uniform k-subsamples. Since ρ follows a hypergeometric distribution,

$$\mathbb{P}(\rho = r) = \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}}, \qquad r = 0, \dots, k.$$

Conditional on $\rho = r$, only Hoeffding components supported on the common indices contribute to the covariance:

$$\mathrm{Cov}(h(x; Z_{S_1}), h(x; Z_{S_2}) \mid \rho = r) = \sum_{c=1}^{r} \binom{r}{c} \zeta_c(x).$$

Thus,

$$\mathrm{Var}\Big(\hat{F}_{n,k}(x)\Big) = \sum_{r=0}^{k} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} \left[\sum_{c=1}^{r} \binom{r}{c} \zeta_c(x)\right].$$

This expression hides the coefficient of each $\zeta_c(x)$ inside a hypergeometric-binomial convolution over all $r \geq c$. The proof of Theorem 3.1 avoids this reduction by counting the joint inclusion probability of each fixed interaction subset directly.

E.2 Comparison with Asymptotic Projection Analyses

Classical asymptotic analyses often use first-order Hájek projections, which are well suited for normality and variance estimation (Mentch & Hooker, 2016; Wager & Athey, 2018; Xu et al., 2024). In regimes where higher-order terms vanish relative to the first-order component, these methods effectively truncate the Hoeffding spectrum. This is appropriate for inference-oriented asymptotics, but it obscures the finite-sample contribution of higher-order interactions.

Theorem 3.1 instead keeps all interaction orders and assigns each order an explicit attenuation factor $\gamma_c(n, k)$. This distinction is important for high-capacity learners, whose variance may concentrate in high-order components associated with interpolation or local memorization (Wyner et al., 2017).

F Proof for Theorem 3.1

Proof. Let $[n] = \{1, 2, \dots, n\}$ denote the finite index set of the training dataset. We define the discrete subsampling space $\Omega = \{S \subseteq [n] : |S| = k\}$. The subagged estimator $\hat{F}_{n,k}(x)$ induces a uniform probability measure $\mathcal{U}$ over Ω, where $\mathbb{P}_{\mathcal{U}}(S) = \binom{n}{k}^{-1}$ for any realization $S \in \Omega$.

The total variance of $\hat{F}_{n,k}(x)$ is given by the expected covariance of two base learners trained on index sets S_1, S_2 drawn independently from Ω under the measure $\mathcal{U}$. Note that the covariance is evaluated with respect to the unknown data-generating distribution, while the outer expectation is with respect to the algorithmic resampling:

$$\mathrm{Var}(\hat{F}_{n,k}(x)) = \mathbb{E}_{S_1, S_2 \sim \mathcal{U}} \left[\mathrm{Cov}(h(x; Z_{S_1}), h(x; Z_{S_2}))\right]. \tag{11}$$

While Lemma 2.3 establishes the orthogonality of interaction subsets $C \subseteq [k]$ within a single base learner, we now lift this analysis to the entire finite population $[n]$. Consider a specific interaction target subset $C \subseteq [n]$ with $|C| = c$. By the established Hoeffding multi-linear orthogonality, the functional covariance between base learners trained on different subsamples S_1 and S_2 is completely decoupled into independent orthogonal contributions. A specific c-th order component $\zeta_c(x)$ contributes to the covariance if and only if its underlying index set C is fully contained in the intersection of both subsamples, i.e., $C \subseteq (S_1 \cap S_2)$. Using the indicator function $\mathbb{I}_{\{C \subseteq S\}}$, we write:

$$\mathrm{Cov}(h(x; Z_{S_1}), h(x; Z_{S_2})) = \sum_{c=1}^{k} \sum_{\substack{C \subseteq [n] \\ |C| = c}} \zeta_c(x) \mathbb{I}_{\{C \subseteq S_1\}} \mathbb{I}_{\{C \subseteq S_2\}}.$$

Substituting this into Eq. (11). By invoking the linearity of expectation for finite sums and exploiting the unconditional independence of the subsampling events S_1 and S_2, we obtain:

$$\begin{aligned}\mathrm{Var}(\hat{F}_{n,k}(x)) &= \mathbb{E}_{S_1,S_2\sim\mathcal{U}}\left[\sum_{c=1}^{k}\sum_{|C|=c}\zeta_c(x)\mathbb{I}_{\{C\subseteq S_1\}}\mathbb{I}_{\{C\subseteq S_2\}}\right]\\ &= \sum_{c=1}^{k}\zeta_c(x)\sum_{|C|=c}\mathbb{E}_{S_1\sim\mathcal{U}}[\mathbb{I}_{\{C\subseteq S_1\}}]\cdot\mathbb{E}_{S_2\sim\mathcal{U}}[\mathbb{I}_{\{C\subseteq S_2\}}]\\ &= \sum_{c=1}^{k}\zeta_c(x)\sum_{|C|=c}\left(\mathbb{P}_{\mathcal{U}}(C\subseteq S)\right)^2.\end{aligned}$$

For a fixed target subset C of size c, constructing a realization $S \in \Omega$ such that $C \subseteq S$ necessitates fixing the c elements of C and uniformly choosing the remaining $k-c$ elements from the complement set. Thus:

$$\mathbb{P}_{\mathcal{U}}(C\subseteq S) = \frac{|\{S\in\Omega \mid C\subseteq S\}|}{|\Omega|} = \frac{\binom{n-c}{k-c}}{\binom{n}{k}}.$$

Since there are exactly $\binom{n}{c}$ possible target subsets of size c within $[n]$, the inner summation evaluates explicitly as:

$$\begin{aligned}\sum_{|C|=c}\left(\mathbb{P}_{\mathcal{U}}(C\subseteq S)\right)^2 &= \sum_{|C|=c}\left(\frac{\binom{n-c}{k-c}}{\binom{n}{k}}\right)^2\\ &= \binom{n}{c}\left(\frac{\binom{n-c}{k-c}}{\binom{n}{k}}\right)^2\\ &= \left[\frac{\binom{n}{c}\binom{n-c}{k-c}}{\binom{n}{k}}\right]\cdot\frac{\binom{n-c}{k-c}}{\binom{n}{k}}\\ &= \left[\frac{\frac{n!}{c!(n-c)!}\cdot\frac{(n-c)!}{(k-c)!(n-k)!}}{\frac{n!}{k!(n-k)!}}\right]\cdot\left(\frac{\frac{(n-c)!}{(k-c)!(n-k)!}}{\frac{n!}{k!(n-k)!}}\right)\\ &= \left[\frac{n!}{c!(k-c)!(n-k)!}\cdot\frac{k!(n-k)!}{n!}\right]\cdot\left(\frac{(n-c)!\,k!}{(k-c)!\,n!}\right)\\ &= \left[\frac{k!}{c!(k-c)!}\right]\cdot\left(\frac{k(k-1)\cdots(k-c+1)}{n(n-1)\cdots(n-c+1)}\right)\\ &= \binom{k}{c}\prod_{j=0}^{c-1}\frac{k-j}{n-j}.\end{aligned}$$

By substituting this identity back into the variance expansion and defining the attenuation factor as $\gamma_c(n,k) \triangleq \prod_{j=0}^{c-1}\frac{k-j}{n-j}$, we recover the exact decoupling:

$$\mathrm{Var}(\hat{F}_{n,k}(x)) = \sum_{c=1}^{k}\gamma_c(n,k)\binom{k}{c}\zeta_c(x).$$

This completes the proof. □

G Proof of Theorem 3.2

Proof. Let S be a random subsample of size k drawn uniformly without replacement from $[n]$. For a fixed subset $C = \{i_1, \dots, i_c\}$ of size c, let E_j denote the event $\{i_j \in S\}$ for $j = 1, \dots, c$. By the symmetry of the uniform measure $\mathcal{U}$, the attenuation factor is the joint probability:

$$\gamma_c(n, k) = \mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_c).$$

Applying the chain rule of probability:

$$\gamma_c(n, k) = \mathbb{P}(E_1) \prod_{j=1}^{c-1} \mathbb{P}(E_{j+1} \mid E_1, \dots, E_j). \tag{12}$$

The marginal probability is $\mathbb{P}(E_1) = \frac{k}{n} = \alpha$. For $j \geq 1$, the conditional probability is:

$$\mathbb{P}(E_{j+1} \mid E_1, \dots, E_j) = \frac{k-j}{n-j}.$$

To compare this with the marginal ratio α, consider the difference:

$$\begin{aligned} \frac{k-j}{n-j} - \frac{k}{n} &= \frac{n(k-j) - k(n-j)}{n(n-j)} \\ &= \frac{nk - nj - kn + kj}{n(n-j)} \\ &= \frac{j(k-n)}{n(n-j)}. \end{aligned}$$

Under the condition $n > k$, the numerator $j(k-n)$ is strictly negative for all $j \geq 1$. Thus:

$$\mathbb{P}(E_{j+1} \mid E_1, \dots, E_j) < \mathbb{P}(E_1) = \alpha. \tag{13}$$

Substituting Eq. (13) into the chain rule Eq. (12) for $c \geq 2$:

$$\gamma_c(n, k) < \alpha \prod_{j=1}^{c-1} \alpha = \alpha^c.$$

The equality $\gamma_c(n, k) = \alpha^c$ holds if $c = 1$ (single event) or if $n = k$ (where $\mathbb{P}(E_{j+1} \mid \dots) = 1$). For all other cases, $\gamma_c(n, k) < \alpha^c$.

For a fixed interaction order c, we analyze the limit of the product form:

$$\begin{aligned} \lim_{n\to\infty, \frac{k}{n}\to\alpha} \gamma_c(n, k) &= \lim_{n\to\infty} \prod_{j=0}^{c-1} \frac{k-j}{n-j} \\ &= \prod_{j=0}^{c-1} \left(\lim_{n\to\infty} \frac{k-j}{n-j} \right). \end{aligned} \tag{14}$$

Expanding each term in the limit:

$$\begin{aligned} \lim_{n\to\infty, \frac{k}{n}\to\alpha} \frac{k-j}{n-j} &= \lim_{n\to\infty, \frac{k}{n}\to\alpha} \frac{k/n - j/n}{1 - j/n} \\ &= \frac{\lim(k/n) - \lim(j/n)}{1 - \lim(j/n)} \\ &= \frac{\alpha - 0}{1 - 0} = \alpha. \end{aligned} \tag{15}$$

Substituting Eq. (15) back into the product Eq. (14):

$$\lim_{n\to\infty, \frac{k}{n}\to\alpha} \gamma_c(n, k) = \prod_{j=0}^{c-1} \alpha = \alpha^c.$$

This demonstrates that as $n \to \infty$, the negative association between sampling events vanishes, and the attenuation converges to a pure geometric filter. □

H Details for Section 4

H.1 Bias Model and MSE Decomposition

Assumption 4.1 abstracts the standard algebraic bias decay used in nonparametric regression. For estimators of Hölder-smooth regression functions, minimax theory yields polynomial rates determined by smoothness and dimension (Stone, 1982; Györfi et al., 2002). In tree-based models, Biau (2012) establishes algebraic bias decay for a random forest model, with an exponent depending on the number of active features. Such assumptions are also common in asymptotic analyses of random forests (Wager & Athey, 2018).

H.2 Stationary Condition for the Continuous Envelope

A direct first-order analysis requires solving $\partial\mathcal{M}(\alpha;\boldsymbol{\zeta},M)/\partial\alpha = 0$. Let

$$H_c(\alpha) = \alpha^c \frac{\Gamma(\alpha n+1)}{\Gamma(c+1)\Gamma(\alpha n-c+1)}.$$

Using the product representation

$$H_c(\alpha) = \alpha^c \frac{1}{c!}\prod_{j=0}^{c-1}(\alpha n - j),$$

we obtain

$$\frac{\partial}{\partial\alpha}\log H_c(\alpha) = \frac{c}{\alpha} + \sum_{j=0}^{c-1}\frac{n}{\alpha n - j}.$$

Therefore the stationary equation is

$$\begin{aligned} 0 &= \frac{\partial\mathcal{M}}{\partial\alpha} \\ &= -2\beta B_0 n^{-2\beta}\alpha^{-2\beta-1} + \sum_{c=1}^{M}\left(\alpha^c\frac{\Gamma(\alpha n+1)}{\Gamma(c+1)\Gamma(\alpha n-c+1)}\right)\left[\frac{c}{\alpha}+\sum_{j=0}^{c-1}\frac{n}{\alpha n-j}\right]\zeta_c(x). \end{aligned} \tag{16}$$

This equation couples the fractional bias derivative with logarithmic derivatives of the generalized binomial terms, making a closed-form solution for α^* generally unavailable.

H.3 Static Resampling Rules and Strict Monotonicity

Static resampling rules. Theorem 4.5 shows that a fixed resampling ratio such as $\alpha \approx 0.632$ can be suboptimal across model complexities. This value comes from the asymptotic bootstrap inclusion probability (Breiman, 1996b; Efron & Tibshirani, 1997), not from optimizing the finite-sample bias–variance trade-off. In low-complexity regimes, small α gives little variance benefit but increases the nonparametric bias term $\mathcal{O}((\alpha n)^{-2\beta})$, so larger subsamples are preferred. In high-complexity regimes, high-order spectral mass makes smaller α beneficial because α^c strongly attenuates interaction noise (Wyner et al., 2017; Belkin et al., 2019). Thus α should be viewed as a complexity-dependent regularization parameter rather than a universal constant.

Corollary H.1 (Strict Monotonicity for Interior Solutions)**.** *Suppose the conditions of Theorem 4.5 hold. If the optimal subsampling ratios are interior solutions, i.e.,* $\alpha^*_{(1)}, \alpha^*_{(2)} \in (0,1)$*, then* $\alpha^*_{(2)} < \alpha^*_{(1)}$.

Proof. Theorem 4.5 gives $\alpha^*_{(2)} \le \alpha^*_{(1)}$. Since $f(\alpha,\boldsymbol{\zeta}) = -\mathcal{M}(\alpha;\boldsymbol{\zeta},M)$ has strictly decreasing differences in $(\alpha,\boldsymbol{\zeta})$ under the stated spectral shift, strict monotonicity for interior optima follows from the strict comparative-statics result of Edlin & Shannon (1998). □

Corollary H.1 states that, when neither optimizer lies on the boundary, adding high-order variance requires a strictly smaller subsampling ratio.

I PROOF OF LEMMA 4.2

Proof. Let $\mathcal{D}_n \sim P^n$ denote the training dataset and $\mathcal{U}$ denote the uniform probability measure over the discrete subsampling space $\Omega = \{S \subseteq [n] : |S| = k\}$. The subagged estimator is defined as the expectation of the base learner over the measure $\mathcal{U}$:

$$\hat{F}_{n,k}(x) = \mathbb{E}_{S\sim\mathcal{U}}[h(x; Z_S)].$$

The point-wise MSE at a target point x is defined as the $L_2(P^n)$ distance between the subagged estimator and the true target function $f^*(x)$:

$$\text{MSE}(\hat{F}_{n,k}(x)) = \mathbb{E}_{\mathcal{D}_n\sim P^n}\left[\left(\hat{F}_{n,k}(x) - f^*(x)\right)^2\right].$$

To isolate the structural bias from the sampling variability, we perform a standard bias-variance decomposition by centering the estimator around its expectation $\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)]$:

$$\begin{aligned}
\text{MSE}(\hat{F}_{n,k}(x)) &= \mathbb{E}_{\mathcal{D}_n}\left[\left(\hat{F}_{n,k}(x) - \mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] + \mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] - f^*(x)\right)^2\right] \\
&= \mathbb{E}_{\mathcal{D}_n}\left[\left(\hat{F}_{n,k}(x) - \mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)]\right)^2\right] + \left(\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] - f^*(x)\right)^2 \\
&\quad + 2\,\mathbb{E}_{\mathcal{D}_n}\left[\left(\hat{F}_{n,k}(x) - \mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)]\right)\left(\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] - f^*(x)\right)\right].
\end{aligned}$$

Since the term $\left(\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] - f^*(x)\right)$ is deterministic with respect to the data-generating measure P^n, by the linearity of expectation, the cross-term simplifies to:

$$\begin{aligned}
\text{Cross-Term} &= 2\left(\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] - f^*(x)\right)\cdot\left(\mathbb{E}_{\mathcal{D}_n}\left[\hat{F}_{n,k}(x)\right] - \mathbb{E}_{\mathcal{D}_n}\left[\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)]\right]\right) \\
&= 2\left(\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] - f^*(x)\right)\cdot\left(\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] - \mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)]\right) \\
&= 0.
\end{aligned}$$

This leaves the exact additive decomposition into squared bias and ensemble variance:

$$\text{MSE}(\hat{F}_{n,k}(x)) = \underbrace{\left(\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)] - f^*(x)\right)^2}_{\text{Squared Bias}} + \underbrace{\text{Var}_{\mathcal{D}_n}(\hat{F}_{n,k}(x))}_{\text{Ensemble Variance}}.$$

For the **Squared Bias** term, we evaluate the inner expectation $\mathbb{E}_{\mathcal{D}_n}[\hat{F}_{n,k}(x)]$. By the linearity of expectation and the independence of the algorithmic resampling from the data-generating process, we interchange the order of expectations:

$$\mathbb{E}_{\mathcal{D}_n\sim P^n}[\hat{F}_{n,k}(x)] = \mathbb{E}_{\mathcal{D}_n\sim P^n}\left[\mathbb{E}_{S\sim\mathcal{U}}[h(x; Z_S)]\right] = \mathbb{E}_{S\sim\mathcal{U}}\left[\mathbb{E}_{\mathcal{D}_n\sim P^n}[h(x; Z_S)]\right].$$

Because all training samples in $\mathcal{D}_n$ are independent and identically distributed (i.i.d.), the marginal expectation of the base learner trained on any subset Z_S of size $|S| = k$ is invariant and equivalent to the expectation on the first k samples:

$$\mathbb{E}_{\mathcal{D}_n\sim P^n}[h(x; Z_S)] = \mathbb{E}_{Z_{1:k}\sim P^k}[h(x; Z_{1:k})] \triangleq \theta_k(x).$$

Thus, $\mathbb{E}_{S\sim\mathcal{U}}[\theta_k(x)] = \theta_k(x)$. Substituting this result into the bias component, we find that the subagging operator preserves the structural bias of the base learner:

$$\text{Squared Bias}(\hat{F}_{n,k}(x)) = \left(\theta_k(x) - f^*(x)\right)^2.$$

By Assumption 4.1, this term is strictly upper bounded as Squared Bias $\leq B_0 k^{-2\beta}$.

For the **Ensemble Variance** term, we directly invoke the exact finite-sample variance decomposition from Theorem 3.1:

$$\text{Var}_{\mathcal{D}_n}(\hat{F}_{n,k}(x)) = \sum_{c=1}^{k}\gamma_c(n,k)\binom{k}{c}\zeta_c(x).$$

Combining the upper bound on bias with the exact variance identity yields the final finite-sample MSE bound:

$$\text{MSE}(\hat{F}_{n,k}(x)) \leq B_0 k^{-2\beta} + \sum_{c=1}^{k} \gamma_c(n,k) \binom{k}{c} \zeta_c(x).$$

This completes the proof. □

J Proof of Lemma 4.4

Proof. Recall from Lemma 4.2 that the finite-sample MSE for any discrete $k \in \mathbb{Z}^+$ satisfies:

$$MSE(\hat{F}_{n,k}(x)) \leq B_0 k^{-2\beta} + \sum_{c=1}^{k} \gamma_c(n,k) \binom{k}{c} \zeta_c(x).$$

Under Assumption 4.3, the spectrum is strictly truncated at the maximum interaction depth M, i.e., $\zeta_c(x) = 0$ for all $c > M$. Substituting this structural sparsity into the variance summation yields:

$$\sum_{c=1}^{k} \gamma_c(n,k) \binom{k}{c} \zeta_c(x) = \sum_{c=1}^{\min(k,M)} \gamma_c(n,k) \binom{k}{c} \zeta_c(x).$$

To construct a continuous surrogate envelope, we define the subsampling ratio $\alpha = k/n$. Constraining the domain to $\alpha \in [M/n, 1]$ directly establishes the lower bound $\alpha n \geq M$, which simplifies the truncated limit: $\min(\alpha n, M) = M$. Substituting $k = \alpha n$ into the MSE bound:

$$MSE(\hat{F}_{n,\alpha n}(x)) \leq B_0 (\alpha n)^{-2\beta} + \sum_{c=1}^{M} \gamma_c(n, \alpha n) \binom{\alpha n}{c} \zeta_c(x).$$

Applying the spectral filtering bound $\gamma_c(n, \alpha n) \leq \alpha^c$ from Theorem 3.2:

$$MSE(\hat{F}_{n,\alpha n}(x)) \leq B_0 (\alpha n)^{-2\beta} + \sum_{c=1}^{M} \alpha^c \binom{\alpha n}{c} \zeta_c(x). \tag{17}$$

To establish differentiability over the continuum $\mathbb{R}^+$, we extend the discrete binomial coefficient via the falling factorial expansion:

$$\binom{\alpha n}{c} = \frac{1}{c!} \prod_{j=0}^{c-1} (\alpha n - j) = \frac{(\alpha n)(\alpha n - 1) \cdots (\alpha n - c + 1)}{c!}.$$

Multiplying the numerator and denominator by the Gamma function $\Gamma(\alpha n - c + 1)$ yields:

$$\binom{\alpha n}{c} = \frac{(\alpha n)(\alpha n - 1) \cdots (\alpha n - c + 1) \cdot \Gamma(\alpha n - c + 1)}{c! \cdot \Gamma(\alpha n - c + 1)}.$$

Applying the functional equation $\Gamma(z+1) = z\Gamma(z)$ recursively absorbs the falling factorial:

$$\begin{aligned}
(\alpha n - c + 1)\Gamma(\alpha n - c + 1) &= \Gamma(\alpha n - c + 2) \\
(\alpha n - c + 2)\Gamma(\alpha n - c + 2) &= \Gamma(\alpha n - c + 3) \\
&\vdots \\
(\alpha n)\Gamma(\alpha n) &= \Gamma(\alpha n + 1).
\end{aligned}$$

Substituting this absorbed numerator and noting $\Gamma(c+1) = c!$, we obtain the analytic continuation:

$$\binom{\alpha n}{c} = \frac{\Gamma(\alpha n + 1)}{\Gamma(c+1)\Gamma(\alpha n - c + 1)}. \tag{18}$$

To guarantee that Eq. (18) is globally analytic and strictly avoids the singular poles of the Gamma function at non-positive integers ($\{0, -1, -2, \dots\}$), we evaluate the algebraic lower bound of the denominator's argument. Given $c \le M$ and $\alpha n \ge M$:

$$\alpha n - c + 1 \ge M - M + 1 = 1 > 0.$$

Because the argument is strictly positive, $\Gamma(\alpha n - c + 1) \in \mathbb{R}^+$ uniformly across the continuous domain, verifying the absence of singularities.

Substituting the continuous mapping Eq. (18) into Eq. (17) defines the exact upper-bounding differentiable surrogate:

$$MSE(\hat{F}_{n,\alpha n}(x)) \le B_0(\alpha n)^{-2\beta} + \sum_{c=1}^{M} \alpha^c \frac{\Gamma(\alpha n + 1)}{\Gamma(c+1)\Gamma(\alpha n - c + 1)} \zeta_c(x) \triangleq \mathcal{M}(\alpha; \boldsymbol{\zeta}, M).$$

This completes the proof. □

K Proof of Theorem 4.5

Proof. Let $\mathcal{H}$ be the space of valid Hoeffding variance spectra. For any two algorithms with spectra $\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)} \in \mathcal{H}$, we define the partial order $\succeq$ such that $\boldsymbol{\zeta}^{(2)} \succeq \boldsymbol{\zeta}^{(1)}$ if there exists a critical interaction order c^* satisfying:

$$\zeta_c^{(2)} - \zeta_c^{(1)} = 0, \quad \forall c < c^*,$$
$$\zeta_c^{(2)} - \zeta_c^{(1)} \ge 0, \quad \forall c \ge c^*,$$

with strict inequality $\zeta_c^{(2)} - \zeta_c^{(1)} > 0$ for at least one $c \ge c^*$.

The continuous asymptotic MSE envelope is given by:

$$\mathcal{M}(\alpha; \boldsymbol{\zeta}) = B_0 n^{-2\beta} \alpha^{-2\beta} + \sum_{c=1}^{n} \alpha^c \binom{\alpha n}{c} \zeta_c.$$

To apply the standard maximization form of Topkis's Theorem (Topkis, 1978; 1998), we define the objective function as $f(\alpha, \boldsymbol{\zeta}) = -\mathcal{M}(\alpha; \boldsymbol{\zeta})$. The optimal subsampling ratio satisfies:

$$\alpha^*(\boldsymbol{\zeta}) = \arg \max_{\alpha \in (0,1]} f(\alpha, \boldsymbol{\zeta}).$$

Let $H_c(\alpha)$ denote the variance inflation multiplier for the c-th order spectrum:

$$H_c(\alpha) \triangleq \alpha^c \binom{\alpha n}{c} = \frac{\alpha^c}{c!} \prod_{j=0}^{c-1} (\alpha n - j).$$

We evaluate the decreasing differences of $f(\alpha, \boldsymbol{\zeta})$ with respect to α and $\boldsymbol{\zeta}$. Taking the partial derivative of f with respect to α:

$$\frac{\partial f}{\partial \alpha}(\alpha, \boldsymbol{\zeta}) = 2\beta B_0 n^{-2\beta} \alpha^{-2\beta-1} - \sum_{c=1}^{n} \zeta_c H_c'(\alpha).$$

Evaluating the difference in marginal returns as the spectrum shifts from $\boldsymbol{\zeta}^{(1)}$ to $\boldsymbol{\zeta}^{(2)}$ (where $\boldsymbol{\zeta}^{(2)} \succeq \boldsymbol{\zeta}^{(1)}$), the bias terms perfectly cancel out:

$$\begin{aligned}
\frac{\partial f}{\partial \alpha}(\alpha, \boldsymbol{\zeta}^{(2)}) - \frac{\partial f}{\partial \alpha}(\alpha, \boldsymbol{\zeta}^{(1)}) &= \left(2\beta B_0 n^{-2\beta} \alpha^{-2\beta-1} - \sum_{c=1}^{n} \zeta_c^{(2)} H_c'(\alpha) \right) \\
&\quad - \left(2\beta B_0 n^{-2\beta} \alpha^{-2\beta-1} - \sum_{c=1}^{n} \zeta_c^{(1)} H_c'(\alpha) \right) \\
&= - \sum_{c=1}^{n} \left(\zeta_c^{(2)} - \zeta_c^{(1)} \right) H_c'(\alpha) \\
&= - \sum_{c=c^*}^{n} \left(\zeta_c^{(2)} - \zeta_c^{(1)} \right) H_c'(\alpha).
\end{aligned}$$

To determine the sign of $H_c'(\alpha)$, we compute the first derivative of $H_c(\alpha)$ via logarithmic differentiation. The natural logarithm of $H_c(\alpha)$ is:

$$\ln H_c(\alpha) = c \ln \alpha - \ln(c!) + \sum_{j=0}^{c-1} \ln(\alpha n - j).$$

Differentiating with respect to α:

$$\begin{aligned}\frac{d}{d\alpha} \ln H_c(\alpha) &= \frac{d}{d\alpha}(c \ln \alpha) - \frac{d}{d\alpha} \ln(c!) + \sum_{j=0}^{c-1} \frac{d}{d\alpha} \ln(\alpha n - j) \\ &= \frac{c}{\alpha} + 0 + \sum_{j=0}^{c-1} \frac{n}{\alpha n - j}.\end{aligned}$$

Thus, the derivative $H_c'(\alpha)$ is expanded as:

$$\begin{aligned}H_c'(\alpha) &= H_c(\alpha) \frac{d \ln H_c(\alpha)}{d\alpha} \\ &= \left[\frac{\alpha^c}{c!} \prod_{j=0}^{c-1} (\alpha n - j) \right] \left(\frac{c}{\alpha} + \sum_{j=0}^{c-1} \frac{n}{\alpha n - j} \right).\end{aligned}$$

In the operative subsampling regime, $k = \alpha n \geq c$. Therefore, for all valid $j \in \{0, 1, \ldots, c-1\}$, we have:

$$\alpha n - j \geq \alpha n - (c-1) > 0.$$

Since $\alpha \in (0, 1]$ and $c \geq 1$, we strictly guarantee:

$$\frac{\alpha^c}{c!} \prod_{j=0}^{c-1} (\alpha n - j) > 0, \quad \frac{c}{\alpha} > 0, \quad \text{and} \quad \frac{n}{\alpha n - j} > 0.$$

Consequently, for all $c \geq 1$:

$$H_c'(\alpha) > 0.$$

Given $\boldsymbol{\zeta}^{(2)} \succeq \boldsymbol{\zeta}^{(1)}$, we have $\zeta_c^{(2)} - \zeta_c^{(1)} \geq 0$ for all $c \geq c^*$, with strict inequality $\zeta_c^{(2)} - \zeta_c^{(1)} > 0$ for at least one $c \geq c^*$. It follows that:

$$\sum_{c=c^*}^{n} \left(\zeta_c^{(2)} - \zeta_c^{(1)} \right) H_c'(\alpha) > 0.$$

Therefore, the marginal return difference is strictly negative:

$$\frac{\partial f}{\partial \alpha}(\alpha, \boldsymbol{\zeta}^{(2)}) - \frac{\partial f}{\partial \alpha}(\alpha, \boldsymbol{\zeta}^{(1)}) < 0.$$

This establishes that $f(\alpha, \boldsymbol{\zeta})$ satisfies strictly decreasing differences in $(\alpha, \boldsymbol{\zeta})$, making it strictly submodular in $(\alpha, \boldsymbol{\zeta})$.

By Topkis's Theorem (Topkis, 1978; 1998), the submodularity implies that the set of optimal solutions $\alpha^*(\boldsymbol{\zeta})$ is monotonically non-increasing in $\boldsymbol{\zeta}$ with respect to the strong set order, yielding

$$\alpha^*_{(2)} \leq \alpha^*_{(1)}.$$

This concludes the proof. □

L Bi-Modal Spectrum Case Study

This section provides a stylized example illustrating how the spectral filtering law controls the optimal subsampling ratio. Consider a learner whose Hoeffding spectrum has two nonzero components: a low-order signal component at $c = 1$ and a high-order noise component at $c = M$, with $1 \ll M \leq k$:

$$\zeta_c(x) = \begin{cases} \sigma_1^2 \binom{k}{1}^{-1}, & c = 1, \\ \sigma_M^2 \binom{k}{M}^{-1}, & c = M, \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Here $\sigma_1^2 > 0$ and $\sigma_M^2 > 0$ denote the total variance contributions of the low-order signal and high-order noise. By Lemma 2.5,

$$\mathrm{Var}(h(x; Z_{1:k})) = \binom{k}{1}\zeta_1(x) + \binom{k}{M}\zeta_M(x) = \sigma_1^2 + \sigma_M^2.$$

Using the exact variance identity and the bias model $\mathcal{B}(k)^2 = B_0 k^{-2\beta}$ with $k = \alpha n$, we obtain Eq. (8), where the last line uses Theorem 3.2. This expression shows the basic filtering mechanism: the low-order component is attenuated at rate α, whereas the high-order component is attenuated at approximately α^M.

Ignoring the lower-order correction term, differentiating Eq. (8) gives

$$\frac{\partial \mathcal{M}}{\partial \alpha} = -2\beta B_0 n^{-2\beta}\alpha^{-2\beta-1} + \sigma_1^2 + M\sigma_M^2\alpha^{M-1}. \tag{20}$$

The optimum balances the marginal bias reduction against the marginal variance increase.

Low-complexity regime. When the learner is strongly regularized, the high-order component is absent or weak. The trade-off is then governed mainly by the bias term and the linear variance term $\sigma_1^2\alpha$, so the optimal subsampling ratio is relatively large.

High-complexity regime. When high-order noise dominates, $\sigma_M^2 \gg \sigma_1^2$, the dominant balance in Eq. (20) is

$$2\beta B_0 n^{-2\beta}\alpha^{-2\beta-1} \asymp M\sigma_M^2\alpha^{M-1}.$$

Thus

$$\alpha^* \asymp \left(\frac{2\beta B_0 n^{-2\beta}}{M\sigma_M^2}\right)^{1/(M+2\beta)} = \mathcal{O}\Big((\sigma_M^2)^{-1/(M+2\beta)}\Big).$$

As the high-order variance contribution grows, the optimal subsampling ratio decreases.

Implication for heuristic ratios. The interaction order M controls the sharpness of the filtering path. Because the attenuation is geometric, even a moderate subsampling ratio can strongly suppress very high-order components; for example, $0.63^{100} \approx 10^{-20}$. This helps explain why classical heuristic ratios can work well for some random-forest settings (Breiman, 2001a), while Theorem 4.5 shows that no fixed ratio is uniformly optimal across spectra.

M CGAS Algorithm and Experimental Protocol

M.1 Complexity-Guided Adaptive Subsampling

While Theorem 4.5 characterizes the structural shift in optimal regularization, determining the exact optimizer α^* is analytically and computationally intractable. This computation necessitates access to the unknown nonparametric bias derivative and the complete Hoeffding variance spectrum $\{\zeta_c\}_{c=1}^k$. Although recent advances in infinite-order U-statistics (Wager & Athey, 2018; Xu et al., 2024) permit the empirical estimation of low-order variance components via the infinitesimal jackknife, recovering the entire spectrum incurs prohibitive computational complexity. To operationalize the insights from Theorem 3.2 without requiring explicit spectrum estimation, we introduce CGAS algorithm, an empirical procedure detailed in Algorithm 1.

Rather than estimating the empirical spectrum, CGAS utilizes a structural hyperparameter of the base learner (e.g., maximum tree depth $d_{\max}$) as a proxy for the effective maximum interaction order. By construction, heavily regularized models (e.g., $d_{\max} \leq 10$) constrain the capacity to form multi-sample interactions ($\zeta_c \approx 0$ for large c), whereas unpruned interpolators distribute substantial variance mass to high-order components. Guided by the monotone comparative statics established in Theorem 4.5, CGAS restricts the search grid for α based on this complexity proxy and selects an empirical optimum using internal K-fold cross-validation. K-fold CV is employed instead of Out-of-Bag (OOB) estimation to mitigate the variance instability characteristic of OOB errors when evaluating small pilot ensembles (B_{search}). By conditioning the search space on the structural complexity of the base learner, the CGAS algorithm avoids the suboptimality inherent in adopting fixed

Algorithm 1 Complexity-Guided Adaptive Subsampling (CGAS)

Input: Training data $\mathcal{D}_n$, base learning algorithm h with maximum complexity proxy d_{max} (e.g., max_depth), pilot ensemble size B_{search} (e.g., 30), final ensemble size B_{final} (e.g., 100), number of CV folds K (e.g., 3).
Phase 1: Theoretical Grid Restriction
if d_{max} is unconstrained or $d_{max} > 10$ **then**
  {High-Complexity Regime: Severe high-order interpolation noise detected.}
  {Necessitates stringent geometric filtering to annihilate noise.}
  Initialize search grid: $\mathcal{A} \leftarrow \{0.1, 0.2, 0.3, 0.4\}$.
else
  {Low-Complexity Regime: Variance concentrated in lower-order structural signals.}
  {Requires minimal variance penalty to prevent bias explosion.}
  Initialize search grid: $\mathcal{A} \leftarrow \{0.6, 0.7, 0.8, 0.9, 0.95\}$.
end if
Phase 2: Internal K-Fold Cross-Validation
Initialize $\alpha^* \leftarrow 0.632$ and $MSE_{best} \leftarrow \infty$.
Split $\mathcal{D}_n$ into K strictly disjoint folds.
for each $\alpha \in \mathcal{A}$ **do**
  for $k = 1$ to K **do**
    Train pilot subagged ensemble (size B_{search}) on training folds using ratio α without replacement.
    Compute Mean Squared Error (MSE_k) on the validation fold.
  end for
  Compute $\overline{MSE}_\alpha \leftarrow \frac{1}{K}\sum_{k=1}^{K} MSE_k$.
  if $\overline{MSE}_\alpha < MSE_{best}$ **then**
    $MSE_{best} \leftarrow \overline{MSE}_\alpha$
    $\alpha^* \leftarrow \alpha$
  end if
end for
Phase 3: Final Ensemble Training
Train the final subagged ensemble of size B_{final} on the entire dataset $\mathcal{D}_n$ using the optimal ratio α^*.
Output: The trained final ensemble and the optimal regularization parameter α^*.

resampling heuristics (e.g., $\alpha \approx 0.632$) irrespective of the model's capacity. Consequently, it operationalizes the subsampling ratio as a complexity-dependent tuning parameter rather than a static variance-reduction constant.

M.2 Experimental Protocol

To empirically investigate ***how the optimal subsampling ratio α^* varies in response to the base learner's interaction complexity***, we formulated a controlled experimental setup.

Functional Complexity Regimes We designed two distinct experimental regimes corresponding to the opposite ends of the Hoeffding complexity spectrum:

- **Regime 1: Low-Complexity Smoothers (Shallow Trees on Original Data).** We constructed a low-complexity regime by constraining the base learners to a maximum tree depth of $d_{\max} = 3$ and training them on the unmodified target distributions. Under these structural constraints, the models possess limited capacity to form complex multi-sample interactions, confining their Hoeffding variance spectrum primarily to low-order main effects (ζ_1, ζ_2). Therefore, high-order variance components are negligible, and the bias-variance trade-off is governed predominantly by the nonparametric bias.
- **Regime 2: High-Complexity Interpolators (Deep Trees with Label Noise).** To induce interpolation behavior and amplify high-order variance, we configured the base learners as fully unpruned decision trees ($d_{\max} = \infty$) and injected additive Gaussian noise (scaled to

$1.5\sigma_y$) into the target variable y. In this regime, the trees act as interpolators that memorize the label noise, distributing substantial variance mass across high-order interaction components.

Datasets We evaluated our theoretical framework across a diverse suite of nine public datasets sourced from OpenML, encompassing a wide range of sample sizes and feature dimensions to ensure robust empirical validation. Specifically, the benchmark included *yprop_4_1* (8885 samples, 252 features), *wine_quality* (6497 samples, 12 features), *superconduct* (21263 samples, 82 features), *dataset_sale* (10738 samples, 15 features), *coli_20* (1440 samples, 1025 features), *elevators* (16599 samples, 17 features), *sarcos* (44484 samples, 22 features), *california_housing* (20640 samples, 10 features), and *2dplanes* (40768 samples, 11 features). This diverse collection ensures that our empirical analysis tests the learning dynamics across various data scales, from standard low-dimensional regression tasks to extreme over-parameterized regimes where functional complexity heavily influences the generalization error.

Baselines To strictly ablate the effect of the combinatorial resampling mechanism, all methods were initialized with `random_state=42` and `n_estimators=100`, and benchmark our proposed CGAS against two Subagging variants utilizing universally adopted heuristic sampling fractions without replacement, namely Subagging (0.632) and Subagging (0.8).

Metrics and Evaluation Protocol To guarantee the statistical robustness of our comparative statics, we performed 5×10 repeated K-Fold Cross-Validation (5 repeats of strictly disjoint 10-folds with varying random seeds). We reported the out-of-sample Mean Squared Error (MSE) and Mean Absolute Error (MAE). Furthermore, to rigorously assert the superiority of the dynamically calibrated α^*, we performed a one-sided Wilcoxon Signed-Rank Test across the CV iterations, testing the null hypothesis that the baselines perform equally to or better than CGAS ($p < 0.001$ indicates statistical significance).

Environmental Setting All experiments were conducted locally using Jupyter Notebook in an Anaconda environment on Windows 10 (Version 10.0.19045). The hardware infrastructure utilized an Intel processor (Intel64 Family 6 Model 151 Stepping 2, GenuineIntel) equipped with 20 logical cores. The software stack was built on Python 3.12.4, utilizing the following key libraries: pandas 2.2.2, numpy 1.26.4, scikit-learn 1.7.2, and statsmodels 0.14.2.

N Additional Experiments

N.1 Generalization to Alternative Symmetric Learners

Since the variance identity (Theorem 3.2) applies to any permutation-invariant base learner (Assumption 2.1), we evaluated K-Nearest Neighbors (KNN) to verify that the spectral filtering mechanism operates independently of the specific algorithm. Following the established protocol, we configured KNN into two regimes: in **Regime 1 (Low-Complexity)**, a smoothed KNN ($K = 20$) trained on original data restricts the Hoeffding spectrum to low-order components; in **Regime 2 (High-Complexity)**, an interpolating 1-NN ($K = 1$) trained on noisy targets yields a heavy-tailed variance spectrum driven by high-order multi-sample interactions.

Table 3 details the generalization performance of subagged KNN ensembles, corroborating the comparative statics of Theorem 4.5. For high-complexity 1-NN interpolators, static heuristics ($\alpha \in \{0.632, 0.8\}$) systematically under-regularized the models. By applying a stricter geometric penalty via smaller α, CGAS suppressed high-frequency noise and mitigated severe error inflation, reducing out-of-sample MSE by 20.7% to 27.1% relative to the 0.632 baseline across the datasets (e.g., from 1.1394 to 0.8308 on *california_housing*). Conversely, for low-complexity smoothed KNNs, CGAS avoided nonparametric bias inflation by adaptively favoring larger subsampling ratios to preserve low-order structural signals. It performed comparably or superiorly to fixed baselines, barring a marginal MSE exception on the *wine_quality* dataset (0.6420 vs. 0.6415 for $\alpha = 0.8$). The parallel learning dynamics observed across decision trees and KNN provide empirical evidence that subagging operates as an algorithm-agnostic spectral filter, indicating that the suboptimality of fixed

resampling fractions is a property of finite-sample structural complexity rather than specific model architectures.

Table 3: Out-of-sample MSE and MAE for **KNN** across various datasets under two distinct functional complexity regimes. The best performance metrics in each regime are highlighted in **bold**.

Datasets	Metric	Regime 1 (Low-Complexity)			Regime 2 (High-Complexity)		
		0.632	0.8	CGAS	0.632	0.8	CGAS
yprop_4_1	MSE	0.9649*	0.9660*	**0.9647**	1.3692*	1.5465*	**1.0295**
	MAE	0.6622*	0.6647*	**0.6618**	0.9137*	0.9724*	**0.7900**
wine_quality	MSE	0.6449*	**0.6415**	0.6420	1.1593*	1,2971*	**0.9195**
	MAE	0.6297*	**0.6267**	**0.6267**	0.8543*	0.9031*	**0.7608**
superconduct	MSE	0.1449*	0.1360*	**0.1304**	0.9728*	1.0731*	**0.7637**
	MAE	0.2316*	0.2220*	**0.2157**	0.7863*	0.8244*	**0.6968**
dataset_sale	MSE	0.5321*	0.5273*	**0.5230**	1.1904*	1.3515*	**0.8832**
	MAE	0.4502*	0.4484*	**0.4475**	0.8658*	0.9240*	**0.7404**
coil_20	MSE	0.3312*	0.2932*	**0.2744**	0.9872*	1.1251*	**0.7569**
	MAE	0.3309*	0.2928*	**0.2717**	0.7952*	0.8496*	**0.6946**
elevators	MSE	0.3454*	0.3373*	**0.3321**	1.1401*	1.2973*	**0.8376**
	MAE	0.3715*	0.3679*	**0.3657**	0.8478*	0.9049*	**0.7261**
california_housing	MSE	0.3103*	0.3059*	**0.3036**	1.1394*	1.2981*	**0.8308**
	MAE	0.3862*	0.3820*	**0.3796**	0.8484*	0.9055*	**0.7243**
sarcos	MSE	0.0597*	0.0511*	**0.0460**	1.0263*	1.1686*	**0.7502**
	MAE	0.1634*	0.1505*	**0.1425**	0.78072*	0.8613*	**0.6914**
2dplanes	MSE	0.0713*	0.0729*	**0.0710**	1.0173*	1.1257*	**0.7579**
	MAE	0.2137*	0.2160*	**0.2132**	0.8045*	0.8451*	**0.6944**

Note: All baselines marked with $*$ are significantly different from CGAS according to the Wilcoxon signed-rank test, where * denotes $p < 0.001$.

N.2 Efficacy Against Standard Bootstrap (Random Forests)

To evaluate the practical implications of our framework, we conducted a controlled comparison between the standard Random Forest (RF) (Breiman, 2001a) and a complexity-calibrated variant (RF*). The baseline RF employs default bootstrap sampling (n draws with replacement), whereas RF* preserves the bootstrap mechanism but constrains the subsample size via `max_samples`=α^*, where α^* is determined by the CGAS algorithm.

Although the exact variance decomposition in Theorem 3.1 strictly assumes subagging to preserve multi-linear orthogonality, applying the CGAS-calibrated α^* to bootstrap sampling is theoretically justified. In the strongly regularized regime ($k \ll n$), bagging and subagging are asymptotically contiguous since the probability of duplicate draws vanishes (Bühlmann & Yu, 2002; Wager, 2014). Even outside this asymptotic limit, restricting `max_samples` bounds the expected number of unique instances per tree. This functionally induces a similar geometric attenuation on high-order interactions (ζ_c), allowing α^* to remain an effective proxy for suppressing interpolation noise despite sample replacement.

Figure 2 compares the relative MSE and MAE reductions of RF* against the standard RF. In Regime 1, RF* yielded marginal performance changes (MSE reductions of 0%-2.5%), demonstrating algorithmic safety by avoiding systematic degradation when rapid attenuation is unwarranted. In contrast, in Regime 2, RF* achieved substantial generalization improvements, reducing out-of-sample MSE by 20.4% in *2dplanes* and 13.3% in *superconduct*. These results align with Theorem 3.2. Standard RF bootstrap sampling draws approximately 63.2% unique instances per tree, imposing a fixed expected subsampling fraction of $\alpha \approx 0.632$. While this attenuation profile suffices for the low-complexity Regime 1, it provides insufficient geometric penalty in Regime 2. Furthermore,

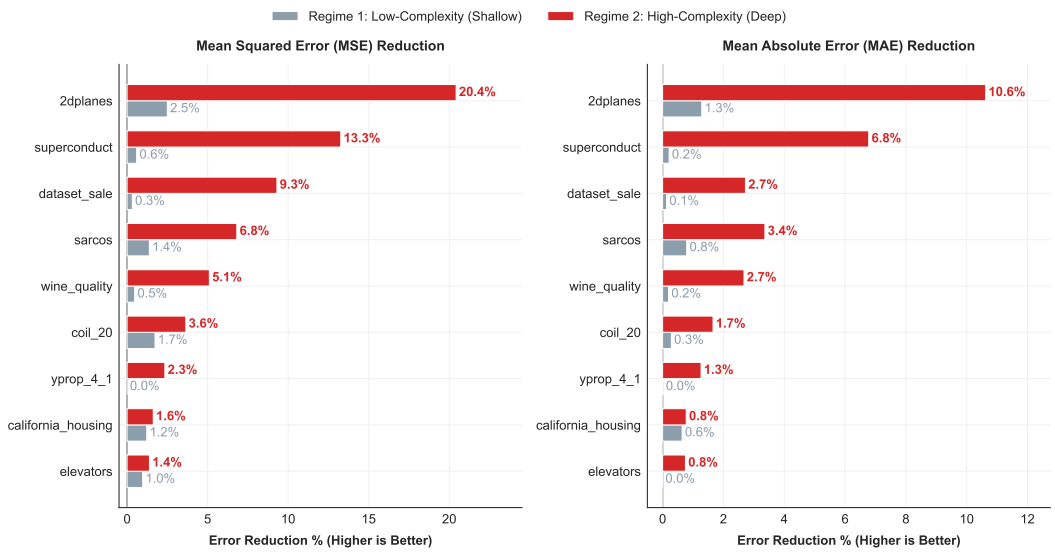

Figure 2: Relative error reduction in out-of-sample MSE (left) and MAE (right) achieved by the theoretically calibrated RF* (bootstrap with `max_samples`$=\alpha^*$) over the standard RF baseline (default bootstrap). The grey bars indicate marginal yet strictly non-negative improvements under the low-complexity regime. The red bars highlight substantial and consistent performance gains under the high-complexity regime, empirically validating that restricting the sampling fraction dynamically suppresses high-order interpolation noise even within a standard bootstrap framework.

the inherent sample duplication ($\sim 36.8\%$) may exacerbate idiosyncratic noise memorization. By restricting the sample size via `max_samples`$=\alpha^*$, RF* enforces a stronger geometric attenuation, effectively controlling high-order variance and improving finite-sample generalization.

O Related Work

O.1 Classical Bias-Variance Analysis and Double Descent

The heuristic that bagging reduces variance without inflating bias originated with Breiman (1996a). Early theoretical formalizations (Bühlmann & Yu, 2002) provided asymptotic justifications by modeling bagging as a smoothing operation over non-differentiable functional components. These continuous approximations, however, inherently rely on the asymptotic regime. The recent success of highly over-parameterized models has renewed interest in the exact mechanics of ensembling, particularly concerning the interpolation regime and double descent phenomena (Wyner et al., 2017; Belkin et al., 2019; Hastie et al., 2022). To define exactly ***which functional components of a model's variance are reduced by resampling, and by what exact proportion***, the analysis must shift from asymptotic smoothing to discrete combinatorial decoupling.

O.2 Infinite-Order U-Statistics and Finite-Sample Bottlenecks

Mathematically, a subagged estimator constructed from a symmetric base learner is equivalent to an incomplete, infinite-order U-statistic (Hoeffding, 1992; Lee, 2019). The Hoeffding-ANOVA decomposition has been pivotal in analyzing Random Forests (Mentch & Hooker, 2016; Wager & Athey, 2018), primarily to establish asymptotic distributional limits (e.g., Gaussianity). Recently, progress has been made in understanding the variance of these estimators. Zhou et al. (2021) explored V-statistic representations to analyze resampling variations, while Xu et al. (2024) utilized the Hoeffding decomposition to establish ratio-consistent variance estimators for large subsample sizes. Furthermore, O'Connell (2026) developed a finite-sample variance identity specific to random forests.

Despite these advances, deriving a generic, finite-sample variance decomposition for *any* symmetric base learner has remained a notoriously difficult combinatorial problem. The mathematical bottleneck stems from the macroscopic evaluation of covariance: the overlap between two random subsamples follows a hypergeometric distribution. Integrating the infinite-order Hoeffding expansions over this hypergeometric distribution yields intractable, nested combinatorial summations. Therefore, existing works either resort to asymptotic limits, which collapse high-order dependencies into first-order projections, or rely on tree-specific structural assumptions (O'Connell, 2026; Xu et al., 2024) to simplify the covariance. A generic functional abstraction providing an exact algebraic identity for the variance reduction remains absent.

O.3 Spectral Filtering in Machine Learning

Analyzing learning algorithms via their spectral properties has precedent in Boolean functions (O'Donnell, 2014) and the implicit bias of neural networks (Rahaman et al., 2019). In ensemble learning, Scornet et al. (2015) and Louppe (2014) explored the structural properties of forests, and Curth et al. (2024) synthesized how tree ensembles act as adaptive smoothers. However, the literature has not yet formalized the combinatorial resampling process itself as a deterministic low-pass filter operating on the Hoeffding spectrum. An explicit algebraic quantification of ***how subagging attenuates high-order variance components***, specifically through isolated transfer functions for each interaction order, remains a theoretical challenge.