

A General Weighting Theory for Ensemble Learning

Beyond Variance Reduction via Spectral and Geometric Structure

Ernest Fokoué

School of Mathematics and Statistics
Rochester Institute of Technology

eMail: epfeqa@rit.edu

Abstract

Ensemble learning is traditionally justified as a variance-reduction device, an explanation that accounts well for unstable base learners such as decision trees [Breiman, 1996, 2001b]. However, this view does not explain the strong empirical performance of ensembles built from intrinsically stable estimators, including splines [Wahba, 1990, Green and Silverman, 1994], kernel ridge regression [Cucker and Smale, 2002], Gaussian process regression [Rasmussen and Williams, 2006], and other smooth function estimators whose variance is already tightly controlled.

In this work, we develop a general weighting theory for ensemble learning that decouples aggregation from randomness and places structure at the center of ensemble design. We formalize ensembles as linear operators acting on a hypothesis space and endow the space of weights with geometric and spectral constraints. Within this framework, we derive a refined bias-variance-approximation decomposition showing how non-uniform structured weights can outperform uniform averaging by simultaneously reducing variance, controlling approximation error, and reshaping the effective hypothesis class.

Our main theorem characterizes conditions under which structured weighting schemes provably dominate uniform ensembles, and shows that optimal weights arise as solutions to constrained quadratic programs. This unified perspective subsumes classical averaging, stacking, and recently proposed Fibonacci-based ensembles as special cases and extends naturally to geometric, sub-exponential, and heavy-tailed weighting laws.

The theory reveals that, for ensembles of low-variance base learners, the principal role of aggregation is not variance reduction, but rather the redistribution of spectral complexity and approximation geometry. Weighted sequences act as geometric operators whose decay properties mediate the trade-off between expressivity and smoothness.

Overall, this work establishes a principled foundation for structure-driven ensemble learning, explaining why ensembles remain effective well beyond the classical high-variance regime and setting the stage for dynamic and distribution-aware weighting schemes developed in subsequent work.

1 Introduction: From Fibonacci Ensembles to a General Weighting Theory

Ensemble learning is one of the most powerful paradigms in modern statistical learning, with landmark developments including bagging [Breiman, 1996], random forests [Breiman, 2001b], boosting [Freund and Schapire, 1997], and stacking [Wolpert, 1992, van der Laan et al., 2007]. The prevailing theoretical justification for ensembles emphasizes *variance reduction*: when base learners are unstable, aggregation stabilizes predictions and improves generalization.

This explanation, while correct and deeply influential, implicitly restricts the scope of ensemble learning to high-variance base learners such as decision trees. By contrast, many of the most classical and mathematically well-understood estimators in statistics—including smoothing splines [Wahba, 1990], penalized regression splines [Green and Silverman, 1994], kernel ridge regression [Cucker and Smale, 2002], Gaussian process regression [Rasmussen and Williams, 2006], and spectral estimators in reproducing kernel Hilbert spaces (RKHS)—are *intrinsically low-variance* due to explicit regularization and spectral shrinkage.

From the classical variance-centric perspective, ensemble methods would therefore appear to have little to offer in such settings: uniform averaging of already stable estimators should yield only marginal gains. Yet empirical and theoretical evidence paints a different picture. Recent results, including the Fibonacci ensembles developed in our companion work, reveal a striking phenomenon: *structured weighting schemes can improve generalization even when variance reduction plays a negligible role* [Wahba, 1990, Poggio and Smale, 2003, De Vito et al., 2006].

This observation motivates the central question of the present paper:

When base learners are smooth, regularized, and low-variance, under what principles can ensemble weighting still improve approximation and generalization?

Our answer proceeds from a change of perspective. In the low-variance regime, the primary role of ensembles is not to suppress noise, but to *reshape the geometry of approximation and the spectral allocation of complexity*. Weighting schemes act as linear operators on ordered dictionaries of functions, reallocating energy across levels of smoothness, frequency, or resolution in a principled manner.

The Fibonacci ensemble provides a canonical example: its weights grow geometrically at a rate governed by the golden ratio, inducing a balance between expressive expansion and spectral stability. In this paper we step beyond this special case to formulate a *General Weighting Theory for Ensemble Learning*, in which Fibonacci weighting becomes one particularly elegant instance within a broad class of admissible weighting laws.

A key structural element of our framework is that many classical function classes arrive with a natural ordering:

- spline bases ordered by smoothness or knot resolution,
- RKHS eigenfunctions ordered by decreasing eigenvalues,
- Fourier or random Fourier features ordered by frequency,
- polynomial bases ordered by degree.

In such settings, weighting sequences interact directly with spectral decay and approximation geometry. The ensemble is no longer an unstructured average, but a *geometrically constrained combination* whose behavior is governed jointly by the dictionary ordering and the decay profile of the weights.

Within this perspective, we introduce a refined decomposition of excess risk that separates classical variance effects from two additional components: *approximation geometry* and *spectral smoothing*. This decomposition explains why certain non-uniform weighting schemes—especially those with controlled geometric decay—can strictly dominate uniform averaging, even when base learners are individually stable.

The present work advances a *general weighting theory for ensemble learning* that recasts aggregation as a structured linear operator acting on a hypothesis space, extending classical aggregation and oracle perspectives [Juditsky et al., 2008, Tsybakov, 2009]. In this view, the choice of weights determines the geometry, spectral properties, and effective approximation power of the ensemble itself. This places ensembles in a conceptual regime where structure and spectral balance, rather than randomness alone, organize their behavior.

Contributions

The contributions of this paper are as follows:

1. We formalize a general class of admissible weighting sequences equipped with geometric and spectral constraints.
2. We develop a refined bias–variance–approximation decomposition tailored to ordered low-variance dictionaries.
3. We derive conditions under which structured weighting provably dominates uniform averaging, and characterize optimal weights via constrained quadratic programs.
4. We connect ensemble weighting to spline smoothing, RKHS regularization, and spectral approximation, thereby unifying ensemble learning with classical estimation theory.

Organization of the Paper

Section 2 introduces the general weighting space and standing assumptions. Section 3 develops the refined bias–variance–approximation decomposition. Section 4 establishes the main results on the superiority of structured weighting schemes. Section 5 discusses implications, extensions, and connections with distribution-adaptive and dynamically evolving weighting strategies that will be explored in subsequent work of this trilogy.

2 The General Weighting Space: Definitions and Assumptions

In this section we formalize the notion of *structured weighting* for ensemble learning. Our goal is to define a general class of admissible weighting schemes that subsumes uniform averaging, Fibonacci weighting, and a wide family of geometric and probabilistic laws, while remaining compatible with stability and generalization guarantees.

We formalize ensemble learning as a problem of weighted aggregation in a Hilbert space. Let $\mathcal{H} \subseteq L^2(P_X)$ be a real separable Hilbert space, and let $h_1, \dots, h_M \in \mathcal{H}$ be base learners obtained from the same training data. An ensemble predictor takes the form

$$\hat{f}_w(x) = \sum_{m=1}^M w_m h_m(x),$$

where $w = (w_1, \dots, w_M)$ is a vector of aggregation weights.

Classical ensemble methods implicitly restrict attention to the *uniform simplex*,

$$\Delta_M = \left\{ w \in \mathbb{R}^M : w_m \geq 0, \sum_{m=1}^M w_m = 1 \right\},$$

leading to simple averaging as in bagging and random forests [Breiman, 1996, 2001a]. More generally, oracle aggregation theory studies data-dependent weights chosen to mimic the best convex combination in hindsight [Juditsky et al., 2008, Tsybakov, 2009].

In this work, we depart from the simplex paradigm and introduce a more general *weighting space* that captures geometric and spectral structure beyond convexity.

2.1 Definition of the Weighting Space

Definition 2.1 (Admissible Weighting Space). Let $\mathcal{W} \subseteq \mathbb{R}^M$ be a closed set of weights. We call \mathcal{W} an *admissible weighting space* if:

(W1) \mathcal{W} is convex and contains the uniform weight vector $w^{\text{unif}} = (1/M, \dots, 1/M)$;

(W2) \mathcal{W} is bounded in ℓ_2 , i.e. $\sup_{w \in \mathcal{W}} \|w\|_2 < \infty$;

(W3) \mathcal{W} is compatible with the geometry of \mathcal{H} , in the sense that $\hat{f}_w \in \mathcal{H}$ for all $w \in \mathcal{W}$.

Conditions (W1)–(W3) are standard in aggregation theory and ensure well-posedness of the ensemble risk minimization problem [Juditsky et al., 2008, Dalalyan and Tsybakov, 2012].

2.2 Risk Decomposition under General Weighting

Let $f^* \in \mathcal{H}$ denote the regression function. The excess risk of the weighted ensemble satisfies

$$\mathbb{E} \left[\|\hat{f}_w - f^*\|_2^2 \right] = \underbrace{\|\mathbb{E}[\hat{f}_w] - f^*\|_2^2}_{\text{bias}} + \underbrace{\mathbb{E} \left[\|\hat{f}_w - \mathbb{E}[\hat{f}_w]\|_2^2 \right]}_{\text{variance}}.$$

This classical decomposition [Geman et al., 1992] implicitly treats the bias term as fixed once the base learners are chosen. However, when w varies over a structured weighting space \mathcal{W} , the bias itself becomes a *design parameter*, reflecting the geometry of the span generated by the weighted learners.

This observation motivates a refinement of the classical bias–variance framework: the approximation properties of the ensemble depend jointly on the choice of base learners and on the admissible weighting geometry. Similar viewpoints appear implicitly in oracle inequalities for aggregation [Tsybakov, 2009], but are not typically emphasized in ensemble design.

2.3 Examples of Weighting Spaces

Uniform simplex. The standard simplex Δ_M corresponds to uniform averaging and classical bagging.

Oracle aggregation weights. Data-dependent weights minimizing empirical risk over Δ_M or related convex sets arise in mirror averaging and exponential weighting schemes [Juditsky et al., 2008, Dalalyan and Tsybakov, 2012].

Structured weighting laws. The weighting spaces introduced in this paper include geometrically constrained sets motivated by spectral decay, stability, and approximation geometry. The Fibonacci weighting scheme studied in Paper I arises as a specific instance of this broader class, illustrating how non-uniform weights can reshape the effective hypothesis space without sacrificing stability.

2.4 Ordered Dictionaries of Base Learners

Let (\mathcal{X}, P_X) be an input space equipped with a probability measure, and let $\mathcal{H} \subseteq L^2(P_X)$ be a real Hilbert space with inner product

$$\langle f, g \rangle = \mathbb{E}[f(X)g(X)].$$

We consider a collection of base learners

$$\mathcal{D}_M = \{h_1, h_2, \dots, h_M\} \subset \mathcal{H},$$

equipped with a *natural ordering* reflecting increasing complexity. Such orderings arise canonically in many classical settings:

- spline bases ordered by degree or knot resolution [Wahba, 1990, Green and Silverman, 1994],
- RKHS eigenfunctions ordered by decreasing eigenvalues [Cucker and Smale, 2002, Steinwart and Christmann, 2008],
- Fourier and random Fourier features ordered by frequency [Rahimi and Recht, 2008],
- polynomial bases ordered by degree.

Throughout, we assume that the ordering is chosen so that h_1 captures the smoothest or lowest-complexity component, while h_M represents the most complex or highest-frequency component available in the dictionary.

2.5 Weighted Ensembles

Given weights $w = (w_1, \dots, w_M)$ with $w_m \geq 0$ and $\sum_{m=1}^M w_m = 1$, we define the corresponding weighted ensemble predictor as

$$\hat{f}_w(x) = \sum_{m=1}^M w_m h_m(x).$$

Uniform averaging corresponds to $w_m = 1/M$, while Fibonacci ensembles arise from geometrically growing weights normalized to sum to one. Our objective is to characterize the general class of weighting sequences for which structured aggregation improves approximation and generalization.

2.6 The Admissible Weighting Space

Definition 2.2 (Admissible Weighting Space). Let \mathcal{W}_M denote the set of all weight vectors $w = (w_1, \dots, w_M)$ satisfying:

(W1) **Nonnegativity and Normalization:**

$$w_m \geq 0, \quad \sum_{m=1}^M w_m = 1.$$

(W2) **Monotone Decay:**

$$w_1 \geq w_2 \geq \dots \geq w_M.$$

(W3) **Square Summability:**

$$\sum_{m=1}^M w_m^2 \leq C_w < \infty,$$

uniformly in M .

Condition (W3) ensures stability and is standard in the analysis of linear aggregation schemes, as it controls the variance contribution of the weights [Bühlmann and Yu, 2003, Koltchinskii, 2011].

2.7 Weighting Families

Within \mathcal{W}_M , several important families arise naturally.

Uniform Weights. The classical choice $w_m = 1/M$ satisfies all conditions but does not exploit the ordering of the dictionary.

Geometric Weights. For $\rho > 1$, define

$$w_m(\rho) = \frac{\rho^m}{\sum_{j=1}^M \rho^j}.$$

These weights emphasize higher-index learners while remaining summable after normalization. Fibonacci weighting corresponds to the minimal geometric growth rate $\rho = \varphi$, the golden ratio.

Sub-Exponential and Polynomial Weights. Weights of the form

$$w_m \propto m^{-\alpha}, \quad \alpha > 1,$$

or

$$w_m \propto \exp(-cm^\beta), \quad 0 < \beta < 1,$$

provide gentler decay and arise naturally in spectral regularization and kernel methods [Caponnetto and De Vito, 2007].

Heavy-Tailed Weights. Distributions such as Zipf or Pareto laws allow slower decay and may be suitable for functions with localized irregularities, though they require stronger control of approximation error to maintain stability.

These families illustrate that Fibonacci weighting is neither arbitrary nor isolated, but occupies a distinguished position at the boundary between expressive expansion and spectral control.

2.8 Standing Assumptions

We now collect the assumptions used throughout the paper.

(A1) Ordered Complexity.

The dictionary $\{h_m\}$ is ordered so that approximation error decreases with m , while spectral complexity (frequency, curvature, or RKHS norm) increases.

(A2) Uniform Variance Control.

There exists $\sigma^2 < \infty$ such that

$$\text{Var}(h_m(X)) \leq \sigma^2 \quad \text{for all } m.$$

(A3) Boundedness.

The learners satisfy $\|h_m\|_\infty \leq B$ almost surely.

(A4) Compatibility with Weighting.

The chosen weighting sequence $w \in \mathcal{W}_M$ respects the ordering of the dictionary, in the sense that higher-complexity learners do not receive larger weights than lower-complexity ones.

Assumptions (A1)–(A4) are mild and satisfied by most classical smoothing and kernel-based estimators. They ensure that weighting interacts with approximation geometry in a controlled manner, without destabilizing the estimator.

2.9 Interpretation

Under this framework, ensemble weighting is no longer viewed as a mere averaging operation, but as a *geometric and spectral operator* acting on an ordered function dictionary. The choice of weights determines how approximation power and smoothness are balanced, independently of classical variance-reduction effects.

This perspective forms the foundation for the refined risk decomposition and generalization theory developed in the sections that follow.

3 A Refined Bias–Variance–Approximation Decomposition

Classical analyses of ensemble learning rely on the bias–variance decomposition to explain the benefits of aggregation. In its traditional form, this framework treats the bias as fixed once the class of base learners is chosen, while the variance is reduced through averaging [Geman et al., 1992, Hastie et al., 2009]. This viewpoint is adequate for highly unstable learners, such as decision trees, but becomes incomplete when the base learners are smooth and intrinsically low-variance.

In this section, we show that when aggregation weights are allowed to vary over a structured weighting space, the ensemble risk admits a refined decomposition in which *approximation geometry* plays a central role. This refinement reveals a mechanism through which ensembles can improve generalization even when variance reduction alone is insufficient.

3.1 Setup and Notation

Let $\mathcal{H} \subseteq L^2(P_X)$ be a real Hilbert space and let $h_1, \dots, h_M \in \mathcal{H}$ be base learners trained on the same data. For a weight vector $w \in \mathcal{W} \subseteq \mathbb{R}^M$, define the ensemble predictor

$$\hat{f}_w = \sum_{m=1}^M w_m h_m.$$

Let $f^* \in \mathcal{H}$ denote the regression function.

We assume that the base learners admit an orthogonalization $\{h_m^\perp\}_{m=1}^M$ in \mathcal{H} , so that

$$\langle h_m^\perp, h_{m'}^\perp \rangle = 0 \quad \text{for } m \neq m'.$$

Such orthogonal decompositions are standard in functional approximation and statistical estimation and play a central role in variance control and Rao–Blackwellization arguments [Lehmann and Casella, 1998].

3.2 Decomposition of the Ensemble Risk

The mean squared error of the ensemble predictor satisfies

$$\mathbb{E}\left[\|\hat{f}_w - f^*\|_2^2\right] = \underbrace{\|\mathbb{E}[\hat{f}_w] - f^*\|_2^2}_{\text{bias}} + \underbrace{\mathbb{E}\left[\|\hat{f}_w - \mathbb{E}[\hat{f}_w]\|_2^2\right]}_{\text{variance}}.$$

When expressed in the orthogonal basis, the variance term simplifies to

$$\text{Var}(\hat{f}_w(X)) = \sum_{m=1}^M w_m^2 \text{Var}(h_m^\perp(X)),$$

revealing an explicit dependence on the squared weights.

The bias term, however, admits a further decomposition. Let

$$\mathcal{H}_w = \text{span}\{w_m h_m^\perp : m = 1, \dots, M\}$$

denote the weighted hypothesis space induced by w . Then

$$\|\mathbb{E}[\hat{f}_w] - f^*\|_2^2 = \underbrace{\|\Pi_{\mathcal{H}_w} f^* - f^*\|_2^2}_{\text{approximation}} + \underbrace{\|\mathbb{E}[\hat{f}_w] - \Pi_{\mathcal{H}_w} f^*\|_2^2}_{\text{estimation bias}},$$

where $\Pi_{\mathcal{H}_w}$ denotes the L^2 -projection onto \mathcal{H}_w .

This decomposition makes explicit a third component, the *approximation error*, which depends on the geometry of the weighted span and varies with the choice of weights.

3.3 Interpretation: Weighting as Geometry Design

The refined decomposition reveals a fundamental principle:

Ensemble learning can improve generalization not only by reducing variance, but by reshaping approximation geometry through structured weighting.

Uniform averaging fixes the geometry of the hypothesis space in advance. In contrast, structured weighting schemes alter the relative contributions of orthogonal components, effectively stretching or compressing directions in \mathcal{H} . This geometric effect allows the ensemble to align more closely with the target function f^* , reducing approximation error without increasing variance.

Related geometric perspectives appear implicitly in oracle inequalities for aggregation [Tsybakov, 2009] and in stability analyses of regularized learning algorithms [Poggio and Smale, 2003], but are rarely articulated as a design principle for ensemble weighting.

3.4 Consequences for Low-Variance Base Learners

For smooth base learners, such as kernel ridge regression, spline estimators, and orthogonal series methods, individual variance is already small and uniform averaging yields diminishing returns. In this regime, the dominant source of error is approximation bias, governed by how well the hypothesis space aligns with the target function [Wahba, 1990, De Vito et al., 2006].

Structured weighting schemes exploit this fact by reallocating weight toward components that contribute most effectively to approximation, while controlling variance through orthogonality and boundedness of \mathcal{W} . This explains why non-uniform ensembles can outperform uniform averaging even for stable learners, a phenomenon observed empirically in Paper I and formalized in the next section.

Classical analyses of ensemble learning decompose the prediction error into bias, variance, and noise components. While this decomposition is effective for high-variance base learners, it obscures the mechanisms by which ensembles improve generalization in regimes where individual learners are already stable.

In this section, we develop a refined decomposition tailored to *ordered, low-variance dictionaries*. The new decomposition isolates the role of weighting in shaping approximation geometry and spectral allocation, thereby explaining why structured ensembles can outperform uniform averaging even when variance reduction is negligible.

3.5 Problem Setup

Let (X, Y) satisfy the regression model

$$Y = f^*(X) + \varepsilon, \quad \mathbb{E}[\varepsilon | X] = 0, \quad \text{Var}(\varepsilon | X) = \sigma^2.$$

Let $\mathcal{H} \subseteq L^2(P_X)$ be a Hilbert space, and let $\{h_1, \dots, h_M\} \subset \mathcal{H}$ be an ordered dictionary of base learners as defined in Section 2. For a weighting vector $w \in \mathcal{W}_M$, define the ensemble estimator

$$\hat{f}_w = \sum_{m=1}^M w_m h_m.$$

We study the excess risk

$$\mathcal{E}(w) = \mathbb{E} \left[\|\hat{f}_w - f^*\|_{L^2(P_X)}^2 \right].$$

3.6 Classical Decomposition and Its Limitations

The standard bias–variance decomposition yields

$$\mathcal{E}(w) = \underbrace{\|\mathbb{E}[\hat{f}_w] - f^*\|_{L^2(P_X)}^2}_{\text{bias}^2} + \underbrace{\mathbb{E} \left[\|\hat{f}_w - \mathbb{E}[\hat{f}_w]\|^2 \right]}_{\text{variance}} + \sigma^2.$$

When each h_m is a regularized estimator (e.g. splines or kernel ridge regression), the variance term is already small and varies little with w . Consequently, this decomposition provides limited insight into why non-uniform weighting schemes can yield systematic improvements.

3.7 Orthogonal Expansion and Approximation Geometry

To expose the effect of weighting, we decompose the dictionary in an orthogonal basis. Let $\{\phi_k\}_{k \geq 1}$ denote an orthonormal basis of \mathcal{H} (e.g. spline basis functions or RKHS eigenfunctions), ordered so that increasing k corresponds to increasing complexity.

Assume that both the target function and the learners admit expansions

$$f^* = \sum_{k \geq 1} \theta_k \phi_k, \quad h_m = \sum_{k \geq 1} a_{m,k} \phi_k.$$

Then the ensemble estimator can be written as

$$\hat{f}_w = \sum_{k \geq 1} \left(\sum_{m=1}^M w_m a_{m,k} \right) \phi_k.$$

The quantity

$$b_k(w) := \sum_{m=1}^M w_m a_{m,k}$$

represents the effective contribution of the k th mode under weighting w .

3.8 The Refined Decomposition

We now decompose the excess risk into three interpretable components.

Theorem 3.1 (Bias–Variance–Approximation Decomposition). *Under assumptions (A1)–(A4), the excess risk admits the decomposition*

$$\mathcal{E}(w) = \underbrace{\sum_{k \geq 1} (b_k(w) - \theta_k)^2}_{\mathcal{A}(w)} + \underbrace{\sum_{k \geq 1} \text{Var}(b_k(w))}_{\mathcal{V}(w)} + \sigma^2,$$

where:

- $\mathcal{A}(w)$ is the approximation geometry term,
- $\mathcal{V}(w)$ is the residual variance term.

Proof. By orthonormality of $\{\phi_k\}$,

$$\|\widehat{f}_w - f^*\|^2 = \sum_{k \geq 1} (b_k(w) - \theta_k)^2.$$

Taking expectation and decomposing each squared term into squared bias plus variance yields the result. \square

3.9 Spectral Smoothing as a Distinct Effect

For low-variance learners, $\mathcal{V}(w)$ is uniformly small and weakly dependent on w . The dominant contribution of weighting therefore appears in $\mathcal{A}(w)$.

We decompose $\mathcal{A}(w)$ further as

$$\mathcal{A}(w) = \underbrace{\sum_{k \leq K(w)} (\theta_k - b_k(w))^2}_{\text{underfitting}} + \underbrace{\sum_{k > K(w)} \theta_k^2}_{\text{unrepresented complexity}},$$

where the effective cutoff $K(w)$ depends on the decay properties of w .

This reveals a third mechanism beyond classical bias and variance:

Definition 3.2 (Spectral Smoothing Term). We define

$$\mathcal{S}(w) = \sum_{k \geq 1} (\theta_k^2 - b_k(w)^2),$$

which measures how weighting redistributes spectral energy across complexity levels.

3.10 Interpretation

The refined decomposition can thus be summarized as

$$\mathcal{E}(w) = \underbrace{\mathcal{A}(w)}_{\text{approximation geometry}} + \underbrace{\mathcal{S}(w)}_{\text{spectral smoothing}} + \underbrace{\mathcal{V}(w)}_{\text{residual variance}} + \sigma^2.$$

In contrast to classical ensemble theory, the dominant effect of structured weighting in the low-variance regime is the joint action of $\mathcal{A}(w)$ and $\mathcal{S}(w)$, which reshape the effective hypothesis space without amplifying noise.

This decomposition explains why geometric and harmonic weighting schemes—such as Fibonacci weighting—can strictly improve generalization even when variance reduction is negligible.

4 Main Theorem: When Structured Weighting Beats Uniform Averaging

We now formalize the intuition developed in the previous section. The theorem below gives sufficient conditions under which a structured weighting scheme strictly improves generalization performance relative to uniform averaging. The key mechanism is a reduction in approximation error without a compensating increase in variance.

4.1 Setting

Let $h_1, \dots, h_M \in \mathcal{H}$ be base learners and let $\hat{f}_w = \sum_{m=1}^M w_m h_m$ denote the ensemble predictor associated with weights $w \in \mathcal{W}$. Let $w^{\text{unif}} = (1/M, \dots, 1/M)$ be the uniform weights and write

$$\hat{f}_{\text{unif}} = \hat{f}_{w^{\text{unif}}}.$$

Let $\mathcal{H}_w = \text{span}\{w_m h_m^\perp\}$ denote the weighted hypothesis space associated with w , and let $\Pi_{\mathcal{H}_w}$ denote the $L^2(P_X)$ projection onto \mathcal{H}_w .

We assume:

- (A1) \mathcal{W} is an admissible weighting space in the sense of Section 2;
- (A2) the orthogonalized components $\{h_m^\perp\}$ satisfy $\text{Var}(h_m^\perp(X)) \leq \sigma^2$ uniformly in m ;
- (A3) the regression function f^* belongs to \mathcal{H} .

Assumption (A2) is natural for stable base learners such as kernel ridge regression or smoothing splines, where the individual estimators are already variance-controlled [Wahba, 1990, De Vito et al., 2006].

4.2 Statement of the Main Theorem

Theorem 4.1 (Structured Weighting Dominance). *Suppose there exists a weight vector $w^* \in \mathcal{W}$ such that*

(C1) (*strict approximation gain*)

$$\|f^* - \Pi_{\mathcal{H}_{w^*}} f^*\|_2^2 < \|f^* - \Pi_{\mathcal{H}_{w^{\text{unif}}}} f^*\|_2^2,$$

(C2) (*controlled variance*)

$$\|w^*\|_2^2 \leq \|w^{\text{unif}}\|_2^2.$$

Then the expected prediction risk of the structured ensemble strictly improves upon uniform averaging:

$$\mathbb{E}\left[\|\hat{f}_{w^*} - f^*\|_2^2\right] < \mathbb{E}\left[\|\hat{f}_{\text{unif}} - f^*\|_2^2\right].$$

4.3 Proof Sketch

By the refined bias–variance–approximation decomposition of Section 3,

$$\mathbb{E}\left[\|\hat{f}_w - f^*\|_2^2\right] = \underbrace{\|f^* - \Pi_{\mathcal{H}_w} f^*\|_2^2}_{\text{approximation}} + \underbrace{\|\mathbb{E}[\hat{f}_w] - \Pi_{\mathcal{H}_w} f^*\|_2^2}_{\text{estimation bias}} + \underbrace{\sum_{m=1}^M w_m^2 \text{Var}(h_m^\perp(X))}_{\text{variance}}.$$

Assumption (A2) implies

$$\sum_{m=1}^M w_m^2 \text{Var}(h_m^\perp(X)) \leq \sigma^2 \|w\|_2^2.$$

Therefore condition (C2) guarantees that the variance of the structured ensemble does not exceed that of the uniform ensemble.

Condition (C1) states that the weighted span \mathcal{H}_w offers a strictly better geometric approximation to f^* than the uniform span. Thus, both approximation error and total risk strictly improve, proving the result.

4.4 Existence of Optimal Weights

The theorem above is existential in nature. Under mild regularity conditions, existence of an optimal weighting vector follows immediately.

Proposition 4.2 (Existence of Optimal Weights). *If \mathcal{W} is compact and convex, there exists*

$$w^{\text{opt}} = \arg \min_{w \in \mathcal{W}} \mathbb{E}\left[\|\hat{f}_w - f^*\|_2^2\right].$$

Moreover, if the risk functional is strictly convex in w , the minimizer is unique.

The proposition follows from standard convex analysis arguments [Rockafellar, 1997]; strict convexity arises naturally when the orthogonalized components are linearly independent.

4.5 Interpretation

The theorem identifies two distinct routes to ensemble improvement:

1. classical variance reduction (as in bagging and random forests);
2. *geometric approximation gain* via structured weighting.

The second mechanism is absent in the traditional bias–variance story, and is precisely the phenomenon exploited by Fibonacci weighting and other structured schemes introduced in this work.

In particular:

Uniform averaging is optimal only when its associated weighted span already provides the best geometric approximation to f^* under the variance constraint.

Otherwise, structured weighting dominates.

In this section we establish the central theoretical result of the paper: for ensembles built from ordered, low-variance dictionaries, uniform averaging is generally *not* optimal. Instead, there exist structured weighting schemes that strictly improve generalization by exploiting approximation geometry and spectral decay.

4.6 Uniform Averaging as a Baseline

Let $w^{\text{unif}} = (1/M, \dots, 1/M)$ denote the uniform weighting, and let \hat{f}_{unif} be the corresponding ensemble estimator.

Uniform averaging ignores the ordering of the dictionary and allocates equal weight to low- and high-complexity components. While this choice is natural and often effective for variance-dominated learners, it fails to exploit the structure present in smooth, ordered dictionaries.

4.7 Existence of Risk-Improving Weighting Schemes

We now state the main theorem.

Theorem 4.3 (Existence of Risk-Improving Structured Weights). *Assume (A1)–(A4). Suppose further that the target function $f^* \in \mathcal{H}$ admits a spectral expansion*

$$f^* = \sum_{k \geq 1} \theta_k \phi_k,$$

with coefficients satisfying

$$|\theta_k| \leq C k^{-\alpha} \quad \text{for some } \alpha > \frac{1}{2}.$$

Then there exists a weighting vector $w^ \in \mathcal{W}_M$ such that*

$$\mathcal{E}(w^*) < \mathcal{E}(w^{\text{unif}}).$$

Moreover, w^ may be chosen to be monotone and geometrically decaying.*

Proof (Sketch). By Theorem 3.1, the excess risk decomposes as

$$\mathcal{E}(w) = \mathcal{A}(w) + \mathcal{V}(w) + \sigma^2.$$

Under assumption (A2), the variance term $\mathcal{V}(w)$ is uniformly bounded and varies weakly across $w \in \mathcal{W}_M$. Consequently, risk differences are dominated by the approximation geometry term $\mathcal{A}(w)$.

For uniform weights, the effective spectral coefficients satisfy

$$b_k(w^{\text{unif}}) = \frac{1}{M} \sum_{m=1}^M a_{m,k},$$

which allocates non-negligible mass to high-frequency modes even when the target coefficients θ_k decay rapidly.

By contrast, consider a geometrically decaying weighting scheme $w_m \propto \rho^m$ with $\rho > 1$. Such weights induce an effective spectral cutoff $K(\rho)$ beyond which $b_k(w)$ decays rapidly. Choosing ρ so that $K(\rho)$ balances the bias incurred by truncation against the decay of θ_k yields

$$\mathcal{A}(w^*) < \mathcal{A}(w^{\text{unif}}).$$

Since $\mathcal{V}(w^*) \approx \mathcal{V}(w^{\text{unif}})$ under the low-variance regime, the strict risk inequality follows. \square

4.8 Near-Optimality of Geometric Weighting

Theorem 4.3 establishes existence but does not yet characterize the structure of optimal weights. The next result shows that geometric weighting is near-optimal in a precise sense.

Theorem 4.4 (Geometric Weights Are Rate-Optimal). *Under the conditions of Theorem 4.3, suppose additionally that the dictionary $\{h_m\}$ resolves spectral modes in increasing order. Then for weights of the form*

$$w_m(\rho) = \frac{\rho^m}{\sum_{j=1}^M \rho^j},$$

there exists $\rho^ > 1$ such that*

$$\mathcal{E}(w(\rho^*)) = \inf_{w \in \mathcal{W}_M} \mathcal{E}(w) + o(1),$$

as $M \rightarrow \infty$.

Proof (Sketch). The geometric decay parameter ρ controls the effective spectral cutoff $K(\rho)$. Matching this cutoff to the decay rate of θ_k yields minimax rates analogous to classical results in spectral regularization and Pinsker theory. The admissibility conditions on \mathcal{W}_M ensure stability. \square

4.9 Fibonacci Weighting as a Distinguished Case

Among geometric weighting schemes, Fibonacci weighting occupies a special position. Its growth rate $\rho = \varphi$ corresponds to the minimal geometric inflation consistent with nontrivial expressive expansion.

Corollary 4.5 (Distinguished Role of Fibonacci Weighting). *Fibonacci weighting achieves a balance between approximation improvement and spectral stability in the sense that it minimizes*

$$\sum_{m=1}^M w_m^2$$

among all geometrically increasing weighting schemes with $\rho > 1$.

This property explains why Fibonacci ensembles often perform competitively with, or better than, more aggressively weighted schemes while remaining numerically stable.

4.10 Interpretation

The results of this section establish a clear and rigorous conclusion:

Uniform averaging is generally suboptimal for ensembles built from ordered, low-variance learners. Structured weighting—particularly geometric and harmonic schemes—can strictly improve generalization by aligning approximation geometry with spectral decay.

This conclusion completes the theoretical arc initiated in Section 2 and Section 3, and provides the foundation for algorithmic and adaptive weighting schemes developed in subsequent work.

5 Consequences, Algorithms, and Outlook

The Main Theorem demonstrates that ensemble improvement need not rely solely on variance reduction. When the base learners are already stable, the dominant mechanism is geometric: structured weighting reshapes the approximation space so that the projection of the target function is closer in $L^2(P_X)$ norm.

This section highlights several consequences of this viewpoint and outlines directions that naturally follow.

5.1 Uniform Averaging as a Special Case

Uniform averaging appears in our framework not as a universal default, but as one specific point in the weighting space \mathcal{W} . The theorem shows that uniform weighting is optimal only in the exceptional case where its associated weighted span already yields the best geometric approximation permitted by the variance constraint.

Thus, the question is no longer

“Should we average?”

but rather

“Which weighting geometry best aligns the ensemble with the target function?”

5.2 Stable Base Learners and the Limits of Variance Reduction

For high-variance learners such as decision trees, uniform aggregation achieves most of its benefit through variance suppression, consistent with classical ensemble theory. However, for stable learners such as kernel ridge regression, splines, and orthogonal series estimators, individual variance is already small and averaging cannot yield substantial improvement.

Our framework explains recent empirical observations that non-uniform weighting can outperform uniform averaging even in this low-variance regime: approximation error, not variance, becomes the dominant quantity, and structured weighting acts directly upon it.

5.3 Spectral and Geometric Perspectives

The dependence of approximation error on the weighted span suggests strong connections to spectral approximation theory, RKHS geometry, and eigenfunction decompositions of kernel operators. Weighting schemes emphasize or suppress components of orthogonal expansions, effectively reshaping the spectrum of the induced estimator.

This opens the door to principled, theoretically justified weighting schemes derived from:

- spectral decay,
- smoothness assumptions on f^* ,
- stability constraints,
- or approximation-theoretic optimality criteria.

The Fibonacci weighting studied in Paper I is one example of such a structured scheme, reflecting a monotone geometric decay motivated by universality and self-similarity properties.

5.4 Algorithmic Implications

The Main Theorem is existential: it asserts that improved weights exist under verifiable conditions. Paper III will address the algorithmic question:

How can optimal or near-optimal weights be found from data?

Possible approaches include:

- convex optimization over \mathcal{W} ,
- entropy-regularized weight learning,
- constrained empirical risk minimization,
- stochastic mirror descent in the weighting space,
- or greedy geometric adaptation of weights.

The refinement of the bias–variance–approximation decomposition developed here will serve as the guiding principle for these algorithms.

5.5 Outlook

The geometric interpretation of ensemble learning developed in this work suggests a broader shift in emphasis:

From randomness to structure;
from variance reduction to approximation design.

This conceptual shift unifies several strands of ensemble methodology and opens new avenues for the principled design of weighting schemes, especially for smooth, low-variance base learners where classical intuition is insufficient.

The computational illustrations and algorithmic developments that realize these ideas in practice are the focus of a companion paper.

The results developed in Sections 2–4 place ensemble learning in a conceptual regime that is distinct from, and complementary to, its classical variance-reduction interpretation. In this section we summarize the principal consequences of the theory, discuss algorithmic implications, and outline directions for future work.

5.6 Conceptual Consequences

The central message of this paper is that ensemble weighting should be viewed as a *geometric and spectral design choice*, rather than merely a device for stabilizing noisy estimators. For ensembles built from ordered, low-variance dictionaries, the dominant effect of weighting lies in its ability to reshape approximation geometry and redistribute spectral complexity.

Several important consequences follow:

- Uniform averaging is generally suboptimal whenever the dictionary admits a meaningful notion of ordered complexity.
- Structured weighting schemes exploit this ordering to achieve a more favorable balance between expressivity and smoothness.
- The benefit of ensembles in this regime persists even when classical variance effects are negligible.

This perspective unifies ensemble learning with classical ideas from spline smoothing, RKHS regularization, and spectral approximation, where the allocation of energy across modes has long been recognized as the key to optimal generalization.

5.7 Algorithmic Implications

Although the present paper is primarily theoretical, the results suggest several practical algorithmic principles.

First, weighting schemes should be chosen in accordance with the ordering of the dictionary. For example, spline bases ordered by knot resolution or RKHS eigenfunctions ordered by eigenvalue naturally invite monotone or geometrically decaying weights.

Second, geometric weighting emerges as a particularly robust and interpretable family. A single decay parameter ρ controls the effective spectral cutoff, making such schemes easy to tune and analyze. Fibonacci weighting appears as a distinguished member of this family, achieving minimal geometric growth while preserving expressive expansion.

Third, the refined decomposition of Section 3 suggests that data-driven selection of weights should target approximation geometry rather than variance alone. This opens the door to adaptive procedures that estimate spectral decay or smoothness directly from the data and select weighting schemes accordingly.

5.8 Why Tree-Based Ensembles Are Not the Focus

A natural question concerns the relationship between the present theory and tree-based ensembles such as Random Forests. While the framework developed here is not incompatible with trees, their dominant source of error is typically variance rather than approximation geometry. As a result, the geometric effects of weighting are largely masked by variance reduction in that setting.

By contrast, smooth estimators—splines, RKHS regressors, and spectral methods—are already stabilized by regularization. It is precisely in this low-variance regime that the role of structured

weighting becomes visible and theoretically meaningful. The present work therefore complements, rather than competes with, existing theories of tree-based ensembles.

5.9 High-Dimensional Considerations

The theory developed here is most transparent in low- and moderate-dimensional settings where ordered dictionaries are readily available. Extending these ideas to high-dimensional problems introduces additional challenges, including the choice of ordering, interactions among features, and the curse of dimensionality.

Nevertheless, many high-dimensional learning problems admit implicit spectral structure—through kernel eigenvalues, neural tangent kernels, or learned feature representations—that may serve as a foundation for structured weighting. The results of this paper suggest that exploiting such structure, rather than relying on uniform aggregation, may be essential for effective ensemble design in complex settings.

5.10 Toward Dynamic and Recursive Weighting Laws

The present work focuses on static weighting schemes. A natural next step is to consider *dynamic* and *recursive* weighting laws, in which ensemble weights evolve over time according to principled update rules. Fibonacci recursions provide one example of such dynamics, but many others are possible.

This perspective motivates the next stage of this research program, in which ensemble learning is viewed as a controlled dynamical system whose stability, expressivity, and generalization properties are governed by the spectral properties of the underlying recursion. These ideas will be developed in a companion paper.

5.11 Closing Remarks

Taken together, the results of this paper suggest a shift in how ensemble methods are conceptualized. Beyond variance reduction, ensembles can be designed to shape approximation geometry and spectral allocation in a deliberate and theoretically grounded manner.

In this sense, ensemble learning becomes less a matter of averaging and more a matter of harmony—balancing growth and restraint, expressivity and stability, in accordance with the intrinsic structure of the function class at hand.

6 Computational Illustrations: When Structure Beats Uniformity

In this section we present simple but informative computational studies designed to illustrate the theoretical results established above. Consistent with the philosophy of this paper, our goal is not to chase benchmark leaderboards, but to demonstrate clearly and transparently the mechanisms through which structured weighting improves generalization.

The experiments are deliberately constructed so that:

- (i) the bias–variance–approximation decomposition is directly observable,

- (ii) both low-variance (e.g. RKHS, splines) and moderate-variance learners are included,
- (iii) the comparison isolates *weighting geometry* rather than model architecture.

6.1 Experimental Setting

We consider the standard nonparametric regression model

$$Y = f_0(X) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

with X sampled uniformly on a compact interval. Two test functions are used:

$$\begin{aligned} f_{\sin}(x) &= \sin(2\pi x), \\ f_{\text{sinc}}(x) &= \begin{cases} \frac{\sin(\pi x)}{\pi x}, & x \neq 0, \\ 1, & x = 0. \end{cases} \end{aligned}$$

These functions were chosen because they embody two qualitatively different approximation regimes: smooth periodicity and localized oscillatory decay.

For each dataset we generate:

$$n_{\text{train}} = 400, \quad n_{\text{test}} = 1000,$$

with Gaussian noise variance σ^2 chosen so that SNR ≈ 5 .

6.2 Base Learners

To highlight that ensembles can improve generalization *even for traditionally low-variance base learners*, we deliberately avoid decision trees and instead use:

1. kernel ridge regression in a Gaussian RKHS,
2. cubic regression splines with fixed knots,
3. polynomial regression models of increasing degree,
4. random Fourier features approximating Gaussian kernels.

These are prototypical examples of learners that already have strong built-in regularization. In classical ensemble folklore, such learners are often considered to have “little to gain” from aggregation. Our results show otherwise: ensembles improve not only by variance reduction, but by reshaping approximation geometry through structured weighting.

6.3 Weighting Schemes Compared

For each family of base learners we construct ensembles under three weighting schemes:

Uniform Averaging.

$$\hat{f}_{\text{unif}} = \frac{1}{M} \sum_{m=1}^M h_m.$$

Fibonacci Weights.

$$\hat{f}_{\text{Fib}} = \sum_{m=1}^M \frac{F_m}{\sum_{j=1}^M F_j} h_m,$$

where (F_m) is the Fibonacci sequence.

Optimal Structured Weights.

We compute the minimizer of the regularized quadratic form

$$\alpha^* = \arg \min_{\alpha \in \mathcal{W}} \left\{ \alpha^\top \Sigma \alpha + \lambda \|\alpha\|_2^2 \right\},$$

where Σ is the empirical covariance matrix of predictions and \mathcal{W} is the structured weight space defined in Section 2.

The third scheme realizes the “oracle” structured weighting discussed in Section 4, while Fibonacci weights serve as a canonical explicit instance of structured geometry without requiring optimization.

6.4 Evaluation Metrics

For each method we compute:

$$\begin{aligned} \text{MSE}_{\text{test}} &= \frac{1}{n_{\text{test}}} \sum (Y^* - \hat{f}(X^*))^2, \\ \text{ISE} &= \int (\hat{f}(x) - f_0(x))^2 dx, \end{aligned}$$

with the integral approximated numerically on a dense grid.

In addition, the bias–variance decomposition is estimated by Monte Carlo replication over $R = 50$ independent training sets:

$$\mathbb{E}[(\hat{f}(x) - f_0(x))^2] = \underbrace{(\mathbb{E}[\hat{f}(x)] - f_0(x))^2}_{\text{bias}^2} + \underbrace{\text{Var}(\hat{f}(x))}_{\text{variance}}.$$

6.5 Representative Figures

The paper includes four representative plots:

1. **Sinusoidal regression with polynomial ensembles**
(uniform vs Fibonacci vs optimal weights).
2. **Sinc regression with polynomial ensembles.**
3. **Sinc regression using random Fourier feature ensembles.**

4. Sine regression using spline ensembles.

In each case we overlay:

true function, training data, three ensemble predictors.

These figures make visually evident that Fibonacci and optimal structured weights adaptively emphasize the right portions of the model family, yielding lower integrated error without increasing estimator variance.

6.6 Summary of Observations

Across all test functions and learner families, the following qualitative phenomena are observed:

- uniform averaging occasionally oversmooths or undersmooths,
- Fibonacci weights substantially reduce integrated squared error,
- optimal structured weights perform best, as predicted theoretically,
- in spline and RKHS settings, improvement occurs *without relying on variance reduction*.

This confirms the main conceptual message of this paper:

Ensembles enhance generalization not only through variance reduction, but also by reorganizing approximation geometry via structured weighting.

7 Conclusion

The classical narrative of ensemble learning emphasizes variance reduction, particularly in the context of unstable base learners such as individual decision trees. In this work we have shown that this narrative, while important, is not complete. Ensembles may improve generalization even when the base learners are already strongly regularized and low-variance — such as spline smoothers, RKHS estimators, kernel methods, or random Fourier feature regressors. The key mechanism is not only variance control, but the reshaping of approximation geometry through structured weighting.

We developed a general framework in which an ensemble is viewed as an element of a *weighting space*. Within this viewpoint, uniform averaging represents only a very special case: it is just one point in a vastly richer geometric object. By imposing mild structural constraints on the admissible weight vectors — Fibonacci structure, monotone majorization, ℓ_2 control, entropy regularization — we showed that the induced hypothesis class changes its approximation behavior in predictable ways. The resulting bias–variance–approximation decomposition makes explicit how weighting geometry redistributes error.

Our main theorem demonstrated the existence of optimal structured weights, strictly outperforming uniform averaging whenever the covariance of the base learner predictions and the approximation residual are suitably aligned. This establishes, in a mathematically transparent manner, that ensembles can improve performance even when variance is not the limiting factor. The computational illustrations confirm the theory: Fibonacci and more general structured weights produce consistent gains across functions and model families, including settings traditionally considered “stable”.

The broader message is conceptual. Ensemble learning need not be understood solely as a device for stabilizing noisy predictors, but as a means for reorganizing approximation power. Weighting is not a cosmetic post-processing step: it is a geometric operator acting on the hypothesis space itself. When the weights are structured rather than uniform, the operator becomes expressive enough to bias learning toward more useful functional subspaces while still being analyzable within statistical learning theory.

This paper is therefore a step toward a more unified perspective on aggregation: *generalization improvement through structured weighting geometry*. The ensuing trilogy continues this development. The present work establishes the static theory; the companion papers explore recursive dynamics and algorithmic instantiations in depth.

This paper has developed a general weighting theory for ensemble learning that extends classical variance-reduction arguments into a broader and more structural regime. By focusing on ensembles built from ordered, low-variance dictionaries, we have shown that aggregation can improve generalization through mechanisms fundamentally different from noise stabilization.

The central insight is that weighting schemes act as geometric and spectral operators on the hypothesis space. When base learners are naturally ordered by complexity—such as spline bases, RKHS eigenfunctions, Fourier features, or polynomial expansions—non-uniform weights reshape approximation geometry and redistribute spectral energy in a principled manner. In this setting, uniform averaging is generally suboptimal.

A refined bias–variance–approximation decomposition revealed that the dominant effects of structured weighting arise from approximation geometry and spectral smoothing rather than classical variance reduction. This perspective explains why geometric and harmonic weighting schemes, including Fibonacci weighting, can yield strict improvements even when individual learners are already stable.

The theory developed here unifies ensemble learning with classical results in spline smoothing, RKHS regularization, and spectral approximation, and places weighting design at the center of ensemble methodology. Rather than treating weights as ad hoc coefficients, we argue that they encode structural laws that govern expressivity, stability, and generalization.

In doing so, this work reframes ensemble learning as a problem of *harmonic design*: balancing growth and restraint, approximation and smoothness, in accordance with the intrinsic structure of the function class under study.

Roadmap to Dynamic and Recursive Weighting Laws. The present work has focused on static weighting schemes, in which ensemble weights are fixed once the dictionary of base learners is specified. A natural and conceptually compelling next step is to allow weights to *evolve* according to principled update rules.

Such dynamic and recursive weighting laws transform ensemble learning into a controlled dynamical system, where stability, expressivity, and generalization are governed by the spectral properties of the underlying recursion. Fibonacci recursions provide a canonical example, but the general theory encompasses a much broader class of second-order and higher-order update mechanisms.

In a companion paper, we develop a theory of recursive ensemble flows, studying their spectral

stability, expressive modes, and learning dynamics. This next stage completes the trilogy by unifying static weighting geometry with temporal recursion, thereby revealing ensemble learning as a structured process of growth with memory rather than a sequence of independent aggregations.

References

- Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. doi: 10.1007/BF00058655.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001a. doi: 10.1023/A:1010933404324.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001b.
- Peter Bühlmann and Bin Yu. Boosting with the L_2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.
- Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
- Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.
- Arnak S. Dalalyan and Alexandre B. Tsybakov. Mirror averaging with sparsity priors. *The Annals of Statistics*, 40(4):2327–2356, 2012.
- Enrico De Vito, Andrea Caponnetto, and Lorenzo Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 6(1):59–83, 2006.
- Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504.
- Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.
- Peter J. Green and Bernard W. Silverman. *Nonparametric Regression and Generalized Linear Models: A Unified Approach*. Chapman and Hall, London, 1994.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. ISBN 978-0387848846.
- Anatoli Juditsky, Philippe Rigollet, and Alexandre B. Tsybakov. Learning by mirror averaging. *The Annals of Statistics*, 36(5):2183–2206, 2008.
- Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Berlin, 2011.
- Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, second edition, 1998.

- Tomaso Poggio and Steve Smale. The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society*, 50(5):537–544, 2003.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184, 2008.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 978-0-262-18253-9.
- R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1997.
- Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.
- Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.
- Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007.
- Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.
- David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.