# Mathematical Trade-offs between Bias and Variance in Ensemble Learning and their influence on Model Generalization in High-Dimensional data

Harsh Agarwal
*Garodia International Centre for Learning Mumbai*

**Abstract- The bias–variance trade-off is a fundamental principle that shapes the generalization ability of machine learning models. In high-dimensional data environments, this trade-off becomes particularly challenging: the sparsity of data inflates variance, while dimensionality reduction or regularization may increase bias. This paper explores the mathematical foundations of bias and variance, presents their decomposition in the context of prediction error, and examines how ensemble learning methods address these challenges. Bagging is shown to reduce variance through decorrelation of base learners, boosting reduces bias by sequentially refining predictions, and stacking combines multiple models to leverage their complementary strengths. The analysis further highlights how ensembles mitigate the curse of dimensionality, where traditional models fail due to distance concentration and instability. Through theoretical discussion and mathematical formulations, the study demonstrates that ensembles provide a balanced approach to error minimization, enabling more robust and reliable generalization in complex, high-dimensional spaces.**

## I. INTRODUCTION

The pursuit of accurate and generalizable predictive models lies at the heart of machine learning. However, achieving this objective is inherently constrained by a fundamental dilemma known as the bias-variance tradeoff. The concept is essentially based on describing the tension between finding a right fit for the model's training data and its capacity to make reliable predictions on unobserved data[1].

Bias essentially refers to the difference between the predicted values of the machine learning model and the true, underlying values of the actual function. It is introduced by the overly simplistic assumptions within the learning algorithm that further fail to identify suitable relationships between data. A model with a high bias directly related to a large magnitude of error in training as well as testing data. For instance, a bias could be represented by using a linear function to plot the points of an exponential equation, leading to a high degree of error between the actual and predicted values.

Variance, contrarily, quantifies the error arising from the model's excessive sensitivity to the complications of the training data. The model is overfitted to exactly follow the sensitive trend of the training data, which also includes its noise, rather than the actual trend. Overfitting occurs because the model seeks to closely replicate the training data, including its minor fluctuations. For example, using a high-degree polynomial might fit the training data perfectly but may perform poorly on unseen data because it captures the noise in the training data as well[2].

Noise, also referred to as irreducible error, represents the intrinsic error that cannot be reduced by any model, regardless of its complexity. This involves the innate randomness, biases, or unmeasurable factors present in the training data itself. Overfitting the model results in replicating this noise, and underfitting the model excludes the noise, as it is insensitive to minor fluctuations.

$$Total\ Error = Bias^2 + Variance + Noise \quad \text{Eqn. 1}$$

Eqn.1 represents the total error in the model.

[1] (GeeksForGeeks, 2025)

[2] (Praveen, 2024)

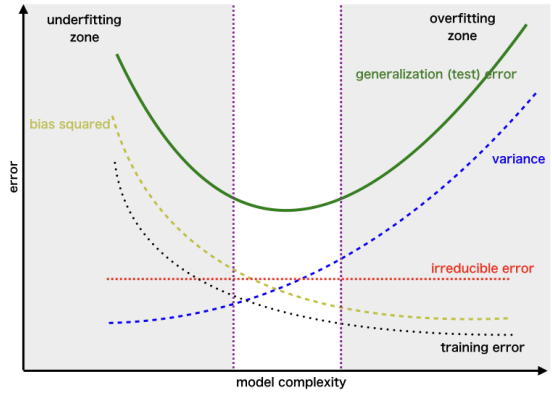


Figure (1)

Figure (1) shows a graphical representation of the relation between error and model complexity. Underfitting refers to the error that is caused due to the ignorance of noise or sensitive data, and overfitting refers to variance, where the model is tuned to follow the sensitive trend of the data, leading to a high degree of error during prediction.

Mathematical Decomposition of Mean Squared Error (MSE)

The relationship between bias, variance, and total error is rigorously formalized through the bias-variance decomposition of the Mean Squared Error (MSE). MSE is a widely recognized and used metric to quantify the average total difference between predicted and actual values. This decomposition is a cornerstone of statistical learning theory, providing a clear framework for analyzing a learning algorithm's expected generalization error.

For an underlying function, $y = f(x) + \varepsilon$, where $\varepsilon$ refers to the irreducible noise in the model with zero mean and variance, $\sigma^2$. The model's predicted values are $\hat{f}(x)$, trained on the dataset, $D$, with MSE at a given point, $x$. The equation can be decomposed as:

$$E_{D,\varepsilon} = (f(x) - E_D[\hat{f}(x)])^2 + (E_D)^2 + \sigma^2 \quad \text{Eqn. } 2^3$$

Where, $E_{D,\varepsilon}$ = Total expected prediction error (MSE);

$(f(x) - E_D[\hat{f}(x)])^2$ = Squared bias, referring to the error due to the model's systematic simplifying assumptions; It is the squared difference between the true function value and the average prediction of the model,

$(E_D)^2$ = Variance, representing the error from the model's sensitivity to the training data,

$\sigma^2$ = Irreducible error, representing the inherent randomness in the model that cannot be accounted for,

The proof of this decomposition includes adding and subtracting the estimator's expected value, $E_D[\hat{f}(x)]$, in the squared error term and then simplifying the expression. The critical step is proving that the cross-product term obtained after this expansion is zero using the properties of expectation.

Although the MSE decomposition is additive and simple, that for the 0-1 loss function (used in classification problems) is more involved and not additively pure. Bias for 0-1 loss in Domingos' unified approach is simply 0 or 1, based on whether the model's "main prediction" (the most common prediction across various training sets) agrees with the actual label. Variance is the probability that the anticipated label is other than this primary prediction. One observation of especial interest for 0-1 loss is that when the bias is very high, adding more variance will sometimes paradoxically reduce the total loss. This is an important distinction in how bias and variance behave according to what loss function is used.

The fundamental tradeoff: Underfitting vs Overfitting

The bias-variance problem is a central and unavoidable problem in supervised machine learning. It illustrates that as the complexity of the model increases, the error should generally decrease because it becomes more sensitive to the training; However, this increased flexibility comes with the cost of a higher variance, also known as overfitting. Conversely, simpler models also exhibit a large bias, or error, due to the inflexibility to sensitive fluctuations. The objective of a model is to, essentially, minimize that error by finding the optimal level of complexity.

## II. ENSEMBLE LEARNING

[3] (Lee, 2025)

Overview of Ensemble Paradigms

Generally, ensemble learning represents a powerful paradigm in machine learning, where the predictions from various individual models are combined to achieve an overall higher generalizability, robustness, and accuracy compared to any single model. Its main goal is to overcome the limitations inherent in individual models by using the underlying premise of collective wisdom. The strength of ensemble learning lies in its ability to smooth out the individual model errors, and capture the broader range of patterns, eventually reducing the noisy data points. This approach is designed to address the bias-variance tradeoff, either by bagging, boosting, or stacking.

Bagging (Bootstrap Aggregating)

Bagging is a parallel ensemble method that includes training multiple base models independently. Each base model is trained on a unique bootstrap sample: random subsets of the original training data created through sampling each replacement. Once trained, their predictions are combined through averaging[4]. Its primary aim is to reduce variance and, consequently, prevent overfitting. It does so because the bootstrap samples inherit a high variance, and averaging them reduces the overall variance.

Boosting (Adaptive and Gradient Boosting)

Unlike bagging, which builds models in parallel, boosting constructs models in a sequential, iterative manner. Each new learner is deliberately shaped by the errors of its predecessor. Specifically, observations that were misclassified or that attracted large residuals in the earlier iterations receive increased weights in the training of the next learner. This weighting mechanism guides the algorithm to devote greater attention to instances that are particularly difficult to classify. When the sequence is complete, the combined prediction is the weighted average of the predictions from all models, with weights determined during training. The driving objective of boosting is to shrink bias and elevate overall predictive performance; to achieve this, the method typically employs simple, high-bias base learners, such as shallow decision trees, often referred to as decision stumps.[5]

Stacking (Stacked Generalization)

Stacking is distinct in that it uses a meta-model to take the outputs from the base learners and combines them. Bagging simply takes the mean or average of the base learners, while boosting takes a weighted sum of the base learners. Stacking takes the predictions from base models and trains another model (the meta-model or "level 1 learner") to learn the best way to combine the predictions made by the base models (the "level 0 learner"). With stacking, you will need to take your training dataset and split it into folds. So, we train the base models on one-fold of data and then make predictions on a fold of data that is “hold-out” data and then train the meta-model on those predictions in the out-of-sample fold. Stacked generalization is just cross-validation, in that it avoids data leakage and ensures the meta-model is learning how to combine predictions that it has never seen before based on the base-learning models. Stacking is viewed as a mechanism to increase overall predictive power; ideally, its performance (the performance of the combination of predictive power) will be better than any one base model. Stacking requires diversity in the base learning models, as the overall, more accurate prediction comes from capturing different relationships in the data.[6]

## III. INFLUENCE ON MODEL GENERALIZATION IN HIGH-DIMENSIONAL DATA

The dimensionality of data significantly impacts the bias-variance trade-off and, consequently, a model's ability to generalize. High-dimensional data presents unique challenges that ensemble methods are particularly well-suited to address.

The Curse of Dimensionality: Challenges for Learning and Generalization

The "Curse of Dimensionality" is a series of phenomena that occur when data is analyzed and structured in high-dimensional spaces that do not appear in lower dimensions. One main effect is that data points get very sparse as dimensions rise. The size

[4] (Cornell, n.d.)
[5] (GeeksForGeeks, 2025)
[6] (Clark & Lee, 2025)

of the feature space increases exponentially with every additional dimension, so the size of data required to have a uniform density of observations or to obtain statistically reliable results also increases exponentially. This sparseness renders it extremely difficult for conventional machine learning methods to find useful patterns and relationships without an impractically large number of training examples. In addition, high-dimensional data sets tend to include many irrelevant or noisy features. These unnecessary features can mask the actual underlying signal, inflate the effective model complexity, and complicate it for the model to generalize well to new data.[7]

$$\lim_{d\to\infty} \frac{max_{i,j}|x_i - x_j| - min_{i,j}|x_i - x_j|}{min_{i,j}|x_i - x_j|} \to 0$$

Eqn. 3

Where, $d$ = dimensionality of the data (number of features),

$x_i, x_j$ = Two data points in the dataset, where each $x$ is a vector of size $d$,

$|x_i - x_j|$ = Euclidian distance between two points (refer to Eqn. 4).

The numerator of Eqn. 3 measures the difference between the farthest (maximum) and the nearest (minimum) points, and dividing by the minimum distance, therefore, normalizes this distance. As the dimensionality $d$ increases, the ratio approaches 0. This means that the minimum and maximum distances become almost nearly equally distant. The formula for the curse of dimensionality highlights why in high-dimensional spaces, distances lose their discriminative power, increasing the difficulty of learning and contributing towards a higher variance in machine learning models.

$$|x_i - x_j| = \sqrt{\sum_{k=1}^{d} (x_{ik} - x_{jk})^2}$$

Eqn. 4: Euclidean distance between two points

Therefore, as the dimensionality increases, models often overfit, due to a higher variance, or they underfit, due to a higher bias. When $d$ increases towards infinity, all points become equidistant, and variance increases. Therefore, in high-dimensionality settings, the overall general error usually increases, as shown in Eqn. 1 and Eqn. 2.

## IV. ROLE OF ENSEMBLE LEARNING TO MITIGATE DIMENSIONALITY CHALLENGES

Ensemble methods are empirically demonstrated to consistently provide a superior bias-variance trade-off in complex and high-dimensional data.

Bagging (e.g., Random Forests): Random Forests are especially reliable and strong in high-dimensional scenarios. They inherently compensate for the high variance of single deep decision trees by averaging their predictions. A central mechanism in Random Forests is random subspace selection, where at every node split, only a random subset of features is examined. This does decorrelate the individual trees and makes the ensemble more robust to irrelevant or noisy features and so controls variance even when lots of features are present. By working in these lower-dimensional subspaces for each tree, Random Forests essentially overcome some of the "curse of dimensionality" effects.

$$VAR_{RF} = \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2$$

Where, $\sigma^2$ = Variance of a single base learning,

$M$ = Number of learners in the ensemble,

$\rho$ = Correlation between individual learners,

Boosting algorithms are very good at learning complex patterns and perform very well in high-dimensional data. By iteratively targeting errors and sequentially minimizing a loss function, they have the ability to learn complex, non-linear patterns that are easily missed by simpler models in high dimensions and thus greatly minimize bias. Implementations like XGBoost and LightGBM are designed specifically for performance and speed while dealing with large-scale and complex datasets, such that they are optimally applicable to high-dimensional problems[8].

[7] (GeeksForGeeks, 2025)

[8] (Clark & Lee, 2025)

Stacking blends different models, taking advantage of the strengths of various learning algorithms, some of which could be better for particular features of high-dimensional data. The heterogeneous blend can result in a stronger and more accurate aggregate prediction in intricate feature spaces by making sense of models that could have varying bias-variance profiles individually, thus smoothing errors and modeling more subtle relationships.

## V. DISCUSSION

The discussion of the bias-variance tradeoff, ensemble learning techniques, and challenges of high-dimensional data demonstrates how all these ideas are fundamentally intertwined. Ultimately, the trade-off quantifies that it is not possible for a single model to minimize both bias and variance simultaneously. Optimizing one tends to cost an increase in the other. This issue is even more critical in higher-dimensional spaces, where sparsity and noise amplify variance and where dimensionality reduction strategies, although beneficial, could introduce bias. A model's generalization capacity is thus immediately influenced by the careful trade-off between these two types of errors.

Ensemble learning offers a systematic approach to resolving this problem. Bagging algorithms like Random Forests reduce variance by averaging predictions over decorrelated base learners and are particularly useful in high-dimensional settings where variance becomes the predominant problem. The boosting methods, conversely, are aimed at bias reduction through iteratively adjusting errors and identifying intricate patterns that naive models tend to miss. Stacking adds a further dimension of versatility through mixing various learner types and leveraging their respective strengths. Collectively, these methods show that ensembles are especially capable of dealing with the bias-variance tradeoff in high-complexity environments.

With their advantages, ensembles have limitations. They tend to need considerable computer power, especially when used in very large or very dimensional data. In addition, their performance relies on ensuring diversity across base learners: when all individual models are very correlated with each other, bagging does not notably decrease variance, and stacking can degenerate into the defects of its building blocks. Such observations point out that although ensembles represent a strong solution, they need to be well-designed and tuned in order to achieve significant improvements in generalization.

In general, the discussion highlights a key understanding: mathematical trade-offs between variance and bias are not just theoretical concepts but concrete principles that have a direct impact on model performance. Ensemble techniques make these principles concrete strategies, providing models that generalize well in high-dimensional data settings where single learners tend to fail.

## VI. CONCLUSION

This study demonstrates that the bias-variance tradeoff provides a fundamental lens through which the performance of machine learning models can be understood, particularly in high-dimensional contexts. As dimensionality increases, models face the dual challenge of higher variance due to distance concentration and potential bias introduced by feature selection or regularization. Ensemble methods such as bagging, boosting, and stacking offer powerful strategies to address these issues by leveraging complementary mechanisms: bagging reduces variance, boosting reduces bias, and stacking combines diverse learners for robustness. Together, these approaches enable ensembles to generalize more effectively than individual models, even under the challenges posed by complex, high-dimensional data. However, the success of ensembles depends on ensuring diversity among base learners and managing computational costs. Overall, ensembles highlight how mathematical insights into bias and variance can be transformed into practical strategies for building reliable, generalizable machine learning models.

## REFERENCE

[1] GeeksForGeeks. (2025, August 6). Bias-Variance Trade Off - Machine Learning. Retrieved August 17, 2025, from https://www.geeksforgeeks.org/machine-learning/ml-bias-variance-trade-off/

[2] Praveen. (2024, August 20). Medium. Retrieved August 17, 2025, from https://praveenkumar2909.medium.com/understa

nding-the-bias-variance-tradeoff-in-machine-learning-f1cb74169250
[3] GeeksForGeeks. (2025, August 6). GeeksForGeeks. Retrieved from https://www.geeksforgeeks.org/machine-learning/ml-bias-variance-trade-off/
[4] Lee, F. (2025, May 21). What is bias-variance tradeoff. Retrieved from https://www.ibm.com/think/topics/bias-variance-tradeoff
[5] Cornell. (n.d.). Bagging. Retrieved from https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote18.html
[6] GeeksForGeeks. (2025, July 23). Ensemble Learning. Retrieved from https://www.geeksforgeeks.org/machine-learning/a-comprehensive-guide-to-ensemble-learning/
[7] Clark, B., & Lee, F. (2025, April 7). What is Gradient Boosting? Retrieved from https://www.ibm.com/think/topics/gradient-boosting
[8] GeeksForGeeks. (2025, July 23). Curse of Dimensionality in Machine Learning. Retrieved from https://www.geeksforgeeks.org/machine-learning/curse-of-dimensionality-in-machine-learning/