A Characterization of Mean Squared Error for Estimator with Bagging

Martin Mihelich
Open Pricer
École Normale Supérieure

Charles Dognin
Verisk | AI, Verisk Analytics

Yan Shu
Walnut Algorithms

Michael Blot
Walnut Algorithms

Abstract

Bagging can significantly improve the generalization performance of unstable machine learning algorithms such as trees or neural networks. Though bagging is now widely used in practice and many empirical studies have explored its behavior, we still know little about the theoretical properties of bagged predictions. In this paper, we theoretically investigate how the bagging method can reduce the Mean Squared Error (MSE) when applied on a statistical estimator. First, we prove that for any estimator, increasing the number of bagged estimators N in the average can only reduce the MSE. This intuitive result, observed empirically and discussed in the literature, has not yet been rigorously proved. Second, we focus on the standard estimator of variance called unbiased sample variance and we develop an exact analytical expression of the MSE for this estimator with bagging. This allows us to rigorously discuss the number of iterations N and the batch size m of the bagging method. From this expression, we state that only if the kurtosis of the distribution is greater than $\frac{3}{2}$, the MSE of the variance estimator can be reduced with bagging. This result is important because it demonstrates that for distribution with low kurtosis, bagging can only deteriorate the performance of a statistical prediction. Finally, we propose a novel general-purpose algorithm to estimate with high precision the variance of a sample.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1 Introduction

Since the popular paper (Breiman, 1996), bootstrap aggregating (or bagging) has become prevalent in machine learning applications. This method considers a number N of samples from the dataset, drawn uniformly with replacement (bootstrapping) and averages the different estimations computed on the different sub-sets in order to obtain a better (bagged) estimator. In this article, we theoretically investigate the ability of bagging to reduce the Mean Squared Error (MSE) of a statistical estimator with an additional focus on the MSE of the variance estimator with bagging.

In a machine learning context, bagging is an ensemble method that is effective at reducing the test error of predictors. Several convincing empirical results highlight this positive effect (Webb and Zheng, 2004; Mao, 1998). However, those observations cannot be generalized to every algorithm and dataset since bagging deteriorates the prediction performance in some cases (Grandvalet, 2004). Unfortunately, it is difficult to understand the reasons explaining why the behavior of bagging differs from one application to another. In fact, only a few articles give theoretical interpretations (Bühlmann et al., 2002; Friedman and Hall, 2007; Buja and Stuetzle, 2016). The difficulty to obtain theoretical results is partially due to the huge number of bootstrap possibilities for the estimator with bagging. For instance, in (Buja and Stuetzle, 2006; Xi Chen and Hall, 2003; Buja and Stuetzle, 2016), authors studied properties of bagging-statistic, which considers all possible bootstrap samples in bagging, and provided a theoretical understanding regarding the relationship between the sample size and the batch size. However, in practice, one cannot consider all possible bootstrap samples. As a result, bagging estimators, which considers a number of randomly constructed samples in bagging are always used. The bagging-statistics studied in the literature (Buja and Stuetzle, 2006; Xi Chen and Hall, 2003; Buja and Stuetzle, 2016) considers the *mean* of bagging samples and the bagging estimator considers the *average* of randomly selected bagging samples, which have more

randomness (We will detail the mathematical difference in section 2.1). These insufficient theoretical guidelines generally lead to arbitrary choices of the fundamental parameters of bagging, such as the batch size m and the number of iterations N.

In this paper, we investigate the theoretical properties of bagging applied to statistical estimators and specifically the impact of bagging on the MSE of those estimators. We provide three core contributions:

- We give a rigorous mathematical demonstration (proof) that the bias of any estimator with bagging is independent from the number of iterations N, while the variance linearly decreases in $\frac{1}{N}$. This intuitive behavior, observed in several papers (Breiman, 1996; Liu et al., 2018), has not yet been demonstrated. We also provide recommendations for the choice of the number of iterations N and the batch size m. We further discuss the implication in a machine learning context. To demonstrate these points, we develop a mathematical framework which enables us, with a symmetric argument, to obtain an exact analytical expression for the bias and variance of any estimator with bagging. This symmetric technique can be used to calculate many other metrics on estimators, we hope that our findings will enable more research in the area.
- We use our framework to further study bagging applied to the standard estimator of variance called unbiased sample variance. This estimator is widely used and a lot of work has been done on specific versions, such as the Jackknife variance estimator (Efron and Stein, 1981) or the variance estimator with replacement (Cho and Cho, 2009). In this paper, we derive an exact analytical formula for the MSE of the variance estimator with bagging. It allows us to provide a simple criteria based on the kurtosis of the sample distribution, which characterizes whether or not the bagging will have a positive impact on the MSE of the variance estimator. We find that on average, applying bagging to the variance estimator reduces the MSE if and only if the kurtosis of the sample distribution is above $\frac{3}{2}$ and the number of bagging iterations N is large enough. From this theorem, we are able to propose a novel, more accurate variance estimation algorithm. This result is particularly interesting since it describes explicit configurations where the application of bagging cannot be beneficial. We further discuss how this result fits into common intuitions in machine learning.
- As a byproduct, using our framework, we provide an alternative proof to the results in (Buja and Stuetzle, 2006) on the bagging-statistics of biased variance estimator. We include this proof and a short comparison of (Buja and Stuetzle, 2006) in the supplementary material.
- Finally, we provide various experiments illustrating and supporting our theoretical results.

The rest of the paper is organized as follows. In Section 2 we present the mathematical framework along with our derivation of the MSE for any bagged estimators. In Section 3, we focus on the particular case of the bagged unbiased sample variance estimator and we provide a new criteria on the variable kurtosis. In Section 4, we support our theoretical findings with three experiments.

2 Bagged Estimators: The More The Better

In this section we rigorously define the notations before presenting the derivation of the different components of the MSE of a bagged statistical estimator, namely the bias and the variance. We subsequently deduce our first contribution: the more iterations N used in the bagging algorithm, the smaller the MSE is, with a linear dependency.

2.1 Notations, Definitions

A dataset, denoted $\ell = (x_i)_{i\in\{1,...n\}}$, is the realization of an independent and identically distributed sample set, noted $L = (X_1, ..., X_n)$, of a variable $X \in \mathcal{X}$. A statistic of the variable X is noted $\theta(X) \in \mathbb{R}$. An estimator of θ, computed from L, is noted $\hat{\theta}(L)$.

By considering $U : \{1, .., m\} \mapsto \{1, .., n\}$, a random function, we denote $L_U = (X_{U(1)}, ..., X_{U(m)})$ a uniform sampling with replacement from L of size m. The random variable U is defined on $\mathcal{U} = \{(u_k)_{k=1\cdots n^m}\}$, the finite set of all m sized sampling with replacement, of cardinal n^m. The bagging method considers N such sampling functions taken uniformly from $\mathcal{U}$, noted $B = (U^1, ..., U^N)$. We can now define the bagging estimator:

$$\tilde{\theta}(L, B) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}(L_{U^i}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}(X_{U^i(1)}, ..., X_{U^i(m)}). \tag{1}$$

Usually, the size of the sample sets is n. Here, we consider the general case where $\hat{\theta}$ is a function of a set of any size m. The number N of iterations is often set to $N \in [|10, 100|]$ (where $[|a, b|]$ represents all the natural numbers between a and b) without rigorous

justification (Breiman, 1996; Dietterich, 2000; Lemmens and Croux, 2006). We discuss this parameter in the following section.

The average with respect to L is noted $\mathbb{E}_L$, and with respect to B, U is noted $\mathbb{E}_B$ and $\mathbb{E}_U$. We assume that $\forall i \in [|1, n^m|]$, $\mathbb{E}_L(\hat{\theta}(L_{U^i})^2) < \infty$. Since B is defined on a finite set, we define $\mathbb{E}_{(L,B)}$ which is the expected value taken with respect to the pair of random variables (L, B) and we have $\mathbb{E}_{(L,B)} = \mathbb{E}_L(\mathbb{E}_B) = \mathbb{E}_B(E_L)$. Using this notation, we remark that the bagging-statistics studied in the literature (Buja and Stuetzle, 2006; Xi Chen and Hall, 2003; Buja and Stuetzle, 2016) is $\mathbb{E}_U(\hat{\theta}(L_U)$, whereas $\tilde{\theta}(L, B)$ defined in equation (1) is an estimator of the bagging-statistics.

There are $(n^m)^N$ possibilities to draw N sampling functions taken uniformly from $\mathcal{U}$ ($(u_1, .., u_N) \in \mathcal{U}^N$) and at L fixed, the average over B takes the form:

$$\mathbb{E}_B(\tilde{\theta}(L, B)) = \frac{1}{(n^m)^N} \sum_{(u_1,..,u_N)\in\mathcal{U}^N} \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}(L_{u_i})$$

2.2 MSE of Estimators with Bagging

In this subsection we state the theorem on the dependence of the bagged estimators in terms of the number of iterations N. The central idea of the proof is to count all the $(n^m)^N$ possible bagged estimators and to use a symmetric argument[1]. The proof of Theorem 2.1 can be found in the supplementary material. We eventually adapt this framework to the case of regression predictors.

Theorem 2.1 *For a general statistic θ, there exists two positive terms F and G, independent of N, such that the MSE of the bagged estimator, $\tilde{\theta}$ defined by* (1)*, satisfies:*

$$\mathrm{MSE}(\tilde{\theta}) = \frac{1}{N}F + G$$

with

$$F = \mathbb{E}_L\big(\mathrm{Var}_U(\hat{\theta}(L_U))\big)$$
$$G = \mathrm{Var}_L\big(\mathbb{E}_U(\hat{\theta}(L_U))\big) + \big(\mathbb{E}_L(\mathbb{E}_U(\hat{\theta}(L_U))) - \theta\big)^2$$

where U is a random variable uniformly distributed on $\mathcal{U}$.

More generally, the following equations hold.

1. $\mathbb{E}_{(L,B)}(\tilde{\theta}(L, B)) = \mathbb{E}_L(\mathbb{E}_U(\hat{\theta}(L_U)))$,
2. $\mathrm{Var}_{(L,B)}(\tilde{\theta}(L, B))$
   $= \frac{1}{N}\mathbb{E}_L\big(\mathrm{Var}_U(\hat{\theta}(L_U))\big) + \mathrm{Var}_L\big(\mathbb{E}_U(\hat{\theta}(L_U))\big)$

[1]For more details, please refer to the complete proof in the supplementary material

Remark 2.1 *$\mathbb{E}_L\big(\mathrm{Var}_U(\hat{\theta}(L_U))\big)$, $\mathrm{Var}_L\big(\mathbb{E}_U(\hat{\theta}(L_U))\big)$ and $\big(\mathbb{E}_L(\mathbb{E}_U(\hat{\theta}(L_U))) - \theta\big)^2$ are positive and do not depend on N. We deduce that:*

1. *The higher the N, the lower the MSE.*
2. *As N goes to ∞, we have:*

$$\lim_{N\to\infty} \mathrm{MSE}(\tilde{\theta})$$
$$= \mathrm{Var}_L\big(\mathbb{E}_U(\hat{\theta}(L_U))\big) + \big(\mathbb{E}_L(\mathbb{E}_U(\hat{\theta}(L_U))) - \theta\big)^2.$$

In this last expression we find the classical MSE decomposition in terms of bias and variance. This expression should be compared with the MSE for the non bagged estimator: $MSE(\hat{\theta}) = \mathrm{Var}_L(\hat{\theta}) + (\mathbb{E}_L(\hat{\theta}) - \theta)^2$. Generally the variance term of the bagged estimator is smaller than the one in the non bagged estimator and it is the opposite for the bias. Nevertheless the sign of the MSE which is the sum of the two terms cannot be easily deduced as we will show later when analysing the variance estimator.

2.3 Bagging for Regression

In the regression setup, the variable X considered is a pair (input, label), $X = (Y, T)$ from $\mathcal{Y} \times \mathcal{T}$. For any input $y \in \mathcal{Y}$, the statistic considered is $\theta_y(Y, T) = \mathbb{E}(T|Y = y)$. The prediction is noted $\hat{\theta}_y(L)$ and the MSE of this estimator represents the prediction error that needs to be reduced. The bagging version of the estimator is noted $\tilde{\theta}_y(L, B)$.

The MSE of the bagged predictor is

$$\mathbb{E}_{y\sim Y}\left(\mathbb{E}_{(L,B)}\left((\tilde{\theta}_y(L, B) - \theta_y)^2\right)\right)$$

According to Theorem 2.1, for every y, there exists two positive values

$$F_y = \mathbb{E}_L\big(\mathrm{Var}_U(\hat{\theta}_y(L_U))\big),$$

and

$$G_y = \mathrm{Var}_L\big(\mathbb{E}_U(\hat{\theta}_y(L_U))\big) + \big(\mathbb{E}_L(\mathbb{E}_U(\hat{\theta}_y(L_U))) - \theta_y\big)^2,$$

independent of N, such that:

$$\mathbb{E}_{(L,B)}\left(\left(\tilde{\theta}_y(L, B) - \theta_y\right)^2\right) = \frac{1}{N}F_y + G_y.$$

Therefore,

$$\mathrm{MSE}(\tilde{\theta}) = \frac{1}{N}\mathbb{E}_{y\sim Y}(F_y) + \mathbb{E}_{y\sim Y}(G_y).$$

We deduce that on average, increasing the number of iterations N can only reduce the average of the MSE of the bagged estimator in a regression setting.

In this section, we demonstrated that the MSE of an estimator with bagging is a linear function of $\frac{1}{N}$ ($= \frac{1}{N}F + G$). The multiplicative component of the dependency, F, is positive. Thus, increasing the number of iterations N can only improve the accuracy of the estimator. This fact has been observed in several empirical studies (Breiman, 1996; Sorokina et al., 2007), but to our knowledge has never been rigorously proved. G represents a lower bound on the MSE of the estimator with bagging. This means that a sufficient N ensures that the term $\frac{1}{N}F$ is negligible compared to G. Moreover, if G is bigger than the MSE of the estimator without bagging, then bagging only deteriorates the precision of the estimator. This deterioration was observed empirically, but the reasons had not yet been theoretically explained (Grandvalet, 2004; Skurichina and Duin, 1998). This result holds for machine learning algorithms in the regression setup as well. In the following section, we apply this general framework to the specific case of the unbiased sample variance estimator.

3 Sample Variance Estimator

In this section, we study the specific case of the bagged unbiased sample variance estimator. Applying the above framework on this particular case, we are able to deduce precise constraints under which using bagging improves on average the MSE of the unbiased sample variance estimator. We derive from this result, given in Theorem 3.3, a criteria expressed in terms of the kurtosis κ of the sample distribution, the batch size m and the number of iterations N. We eventually propose at the end of the section, a novel algorithm which provides a more accurate variance estimation if those criteria are met. Carrying the notation of the precedent section, we assume without loss of generality that $\mathbb{E}(X) = 0$ by taking $X := X - \mathbb{E}(X)$. We denote μ_2 the second moment of the centered random variable and μ_4 the forth moment. Since X is centered, μ_2 is the variance of X and μ_4/μ_2^2 is the kurtosis of X. In the rest of the section $\theta(X) = \mu_2$ is the variance of X and we note $\hat{\theta}(L) = \hat{v}(L) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i)^2$. The version of this estimator with bagging follows definition (1) and is noted $\tilde{v}(L, B)$.

3.1 Bias and Variance of the Sample Variance Estimator with Bagging

The following theorem gives an exact expression of the bias and variance of the variance estimator with bagging.

Theorem 3.1 *Let X be a random variable with finite fourth moment and satisfying $\mathbb{E}(X) = 0$. The bagged variance estimator $\tilde{v}(L, B)$ with batch size m and number of bagging iteration N satisfies:*

1. $\mathbb{E}_{(L,B)}(\tilde{v}(L, B)) = \frac{n-1}{n}\mu_2$,
2. $\mathrm{Var}_{(L,B)}(\tilde{v}(L, B))$
   $= \frac{1}{N}\mathbb{E}_L\big(\mathrm{Var}_U(\hat{v}(L_U))\big) + \mathrm{Var}_L\big(\mathbb{E}_U(\hat{v}(L_U))\big)$;

where

$$\mathbb{E}_L\big(\mathrm{Var}_U(\hat{v}(L_U))\big) \tag{2}$$
$$= \frac{n-1}{nm(m-1)}\left(3m - 3 + \frac{n^2 - 2n + 3}{n^2}(6 - 4m)\right)\mu_2^2$$
$$+ \frac{n-1}{nm(m-1)}\left(m - 1 + \frac{n-1}{n^2}(6 - 4m)\right)\mu_4, \tag{3}$$

and

$$\mathrm{Var}_L\big(\mathbb{E}_U(\hat{v}(L_U))\big) = \frac{(3-n)(n-1)}{n^3}\mu_2^2 + \frac{(n-1)^2}{n^3}\mu_4.$$

Moreover, the MSE of the variance estimator with bagging is:

$$\mathrm{MSE}(\tilde{v}(L, B)) = \mathrm{Var}_{(L,B)}(\tilde{v}(L, B)) + \frac{1}{n^2}\mu_2^2.$$

The proof of Theorem 3.1 can be found in the supplementary material.

3.2 Asymptotic Analysis and Comparison of Estimators

In this section, we analyze the asymptotic behavior of the bagged estimator with respect to the parameters m and N. We also compare the bagged estimator and the non-bagged estimator. Recall that for the standard variance estimator, the MSE is known(see (Rose and Smith, 2002)).

Proposition 3.2 *Assuming that X has a finite fourth moment, the variance of the standard sample variance estimator $\hat{v}(L)$ is given by:*

$$\mathrm{Var}_L(\hat{v}(L)) = \frac{3-n}{n(n-1)}\mu_2^2 + \frac{1}{n}\mu_4.$$

Since this estimator is unbiased, it holds:

$$\mathrm{MSE}(\hat{v}(L)) = \mathrm{Var}_L(\hat{v}(L))$$
$$= \frac{3-n}{n(n-1)}\mu_2^2 + \frac{1}{n}\mu_4.$$

Proposition (3.2) combined with Theorem 3.1 enables us to compare the MSE of $\hat{v}$ and $\tilde{v}(L, B)$ and we find that:

$$\mathrm{MSE}(\tilde{v}(L, B)) - \mathrm{MSE}(\hat{v})$$
$$= \frac{1}{Nm}\left(\mu_4 - \mu_2^2\right) + \frac{1}{n^2}\left(-2\mu_4 + 3\mu_2^2\right) + o\left(\frac{1}{Nm} + \frac{1}{n^2}\right).$$

We state the following theorem.

Theorem 3.3 *As n tends to $+\infty$, bagging reduces on average the MSE of the variance estimator if and only if:*

$$-2\mu_4 + 3\mu_2^2 < 0, \tag{4}$$

and

$$N > \frac{\mu_4 - \mu_2^2}{2\mu_4 - 3\mu_2^2}\frac{n^2}{m}. \tag{5}$$

Therefore, bagging should be used when (4) and (5) are satisfied. Thus N and m should be carefully chosen. Moreover, the gain obtain by using bagging is in $O(\frac{1}{n^2})$. The equation (4) can be rewritten as $\kappa > \frac{3}{2}$, with κ the kurtosis of X.

Remark that if $m \leq n$, then

$$\frac{\mu_4 - \mu_2^2}{2\mu_4 - 3\mu_2^2}\frac{n}{m} > \frac{1}{2}.$$

We deduce that the number of iterations N should be at least $n/2$ in this case. As previously mentioned, the number of iterations N of bagging is often chosen as $N \in [|10, 100|]$ without theoretical justifications (Fazelpour et al., 2016). Theorem 3.3 gives, for the sample variance estimator, a minimum number of iterations N above which bagging improves the estimation.

Note that the condition on the kurtosis ($\kappa > \frac{3}{2}$) is not restrictive. In fact, many classical continuous distributions satisfy this condition. For example, $\kappa = 3$ for a normal distribution, $\kappa = 4.2$ for a logistic distribution, $\kappa = 1.8$ for a uniform distribution, $\kappa = 9$ for an exponential distribution. We now propose a novel algorithm to estimate the variance of a sample.

3.3 Algorithm for Higher Precision Variance Estimation

We deduce from the preceding theoretical results a simple algorithm to estimate the variance of a sample with higher precision. To simplify the procedure, we choose to take the batch size m equal to the full sample size n. Let q, an integer greater or equal to 1, denote a parameter used to control the quality of the resulting estimation. The higher the q, the higher N; and as a result of the Theorem 2.1 and Theorem 3.3, the better the estimation is (at the cost of an increasing computation time). We suggest taking $q \approx 2 - 10$. $\lfloor . \rfloor$ is the floor function. The reader can find an analysis of the algorithmic complexity in the supplementary material.

Algorithm 1 Variance Estimation Algorithm

procedure VARIANCEESTIMATOR($L = (x_1, \cdots, x_n), n, q$)
  $\bar{x} \leftarrow \frac{1}{n}\sum_{i=1}^{n} x_i$
  $\hat{\mu}_4 \leftarrow \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$
  $\hat{v} \leftarrow \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  if $-2\hat{\mu}_4 + 3\hat{v}^2 < 0$ **then**
    $N \leftarrow q \times (\lfloor \frac{\hat{\mu}_4 - \hat{v}^2}{2\hat{\mu}_4 - 3\hat{v}^2} n \rfloor + 1)$
    for $i \in \{1, \cdots, N\}$ **do**
      Draw with replacement n points from the dataset
      $\hat{v}_i \leftarrow \hat{v}(x_{U(1)}, \cdots, x_{U(n)})$
    $\tilde{v} \leftarrow \frac{1}{N}\sum_{i=1}^{N} \hat{v}_i$
    $\hat{v} \leftarrow \tilde{v}$
  return $\hat{v}$

4 Experiments

In this section we present various experiments illustrating and supporting the previous theoretical results. The section begins with experiments on general estimator, with a focus on regression estimators as presented in section 2.3. Then, we present empirical experiments about the bagging version of the sample variance estimator, supporting our theoretical results of section 3. For the experiments, we used Python and specifically the Scikit-Learn (Pedregosa et al., 2011), Numpy (Walt et al., 2011), Scipy (Jones et al., 2001) and Numba (Lam et al., 2015) libraries[2]. More details on the experimental setup can be found in the supplementary material.

4.1 Variance of Bagged Estimators in Regression

In order to empirically validate Theorem 2.1 we performed experiments on two regression predictors: the linear regression and the decision tree regression. In each case, we measured the MSE on the test set after training each model on bagged samples, varying the parameter N of bagging iterations. We generated toy regression datasets using the Scikit-Learn library. Each dataset, has a specific amount of Gaussian noise (σ) applied to the output (0.5 or 5). The number of samples (1000) and the size of the feature space dimension (2) remained constant. To obtain the bagged and non-bagged predictors, we trained the regressors on 5% of the data and tested it on the remaining 95%. This partition was chosen to enable a fast training of the growing number of bagged predictors. We are interested in showing empirically that our theoretical relation holds, not to achieve good absolute results. Using the notations from 2.3: the average $\mathbb{E}_y$ was taken over the test set of size 950. $\mathbb{E}_B$ was taken over 100 trials and $\mathbb{E}_L$ was taken over 10 trials. As expected, we observe that the MSE of those estimators decreases

[2]To reproduce the experiments, please use: `https://github.com/cdcsai/bagging_research`

at a rate of $\frac{1}{N}$. We fit a non-linear function on the resulting datapoints of the form $\hat{y} = \hat{a} + \frac{\hat{b}}{N}$ to better observe this relationship. More details on the experiment setup and hyperparameters used can be found in the supplementary material.

4.2 MSE of Bagged and Non-Bagged Unbiased Variance Estimator

In this experiment, we test the empirical validity of Theorem 3.3. Our condition on the kurtosis comes from the following equality that we proved in the supplementary material: $\mathrm{MSE}(\tilde{v}(L,B)) - \mathrm{MSE}(\hat{v}) = \frac{1}{n^2}\left(-2\mu_4 + 3\mu_2^2\right) + O\left(\frac{1}{Nm} + \frac{1}{n^2}\right)$. Here we set $m = n$. We measured the distance between the MSE of a bagged and non-bagged unbiased sample variance estimator for three classical probability distributions: a standard Gaussian distribution, a uniform distribution on $[-1, 1]$ and a Rademacher distribution. We fixed the number of iterations N to 50, only varying the sample size n, and averaged the estimated variance over 10000 trials. As expected, it follows the same shape as our condition $(-2\mu_4 + 3\mu_2^2)$ divided by n^2.

4.2.1 Kurtosis Condition

In this experiment, we tested the kurtosis condition $\kappa > \frac{3}{2}$. Distributions with kurtosis lower than $\frac{3}{2}$ are unusual as previously mentioned. We designed a distribution whose kurtosis can vary above and below $\frac{3}{2}$ when the parameter p is changing. Let X be a random variable following this distribution. With $p \in [0, 1]$, $a > 0$ and $p + q = 1$:

$$P(X = 1) = P(X = -1) = \frac{p}{2}.$$

and

$$P(X = \sqrt{a}) = P(X = -\sqrt{a}) = \frac{q}{2}.$$

The kurtosis of this distribution is thus:

$$\kappa = \frac{p + qa^2}{(p + qa)^2}.$$

Varying the parameter p between 0 and 1, the kurtosis varies from 1 to ∞ and Figure 3 shows the MSE with and without bagging accordingly. We fixed N to 20, n to 10, a to $\frac{1}{8}$, and we averaged over 100000 trials. As expected, The MSE of the estimator with bagging becomes better than the MSE of the estimator without bagging when the $\frac{3}{2}$ threshold is passed.

5 Conclusion

In this paper, we theoretically investigate the MSE of estimators with bagging. We prove the existence of a linear dependency of the MSE on $\frac{1}{N}$, N being the number of bagged averaged estimators. This dependency enables us to provide guidelines for setting the parameter N. We also explain why bagging is detrimental in some cases. We use the mathematical framework developed in Section 2 to describe the MSE for the sample variance estimator with bagging. It appears that the bagging can reduce the MSE of this estimator if and only if the kurtosis of the distribution is above $\frac{3}{2}$. This condition holds for a large number of classical probability distributions. Using this condition, we eventually propose a novel algorithm for more accurate variance estimation. We hope that our mathematical framework will help generalize our results to more complex estimators like decision trees or neural networks.

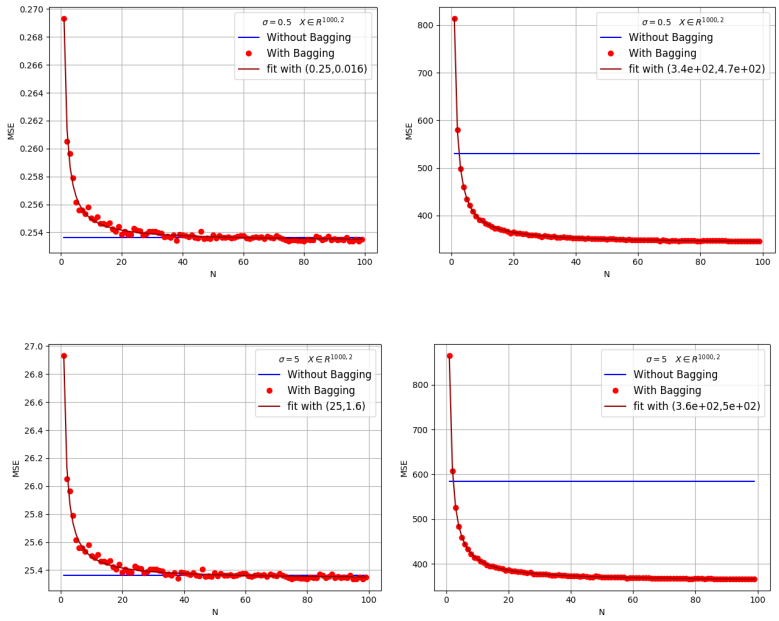
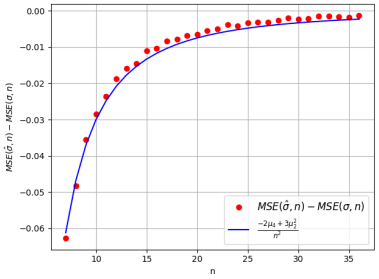
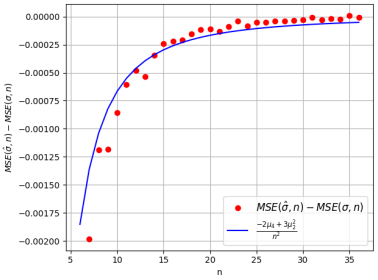


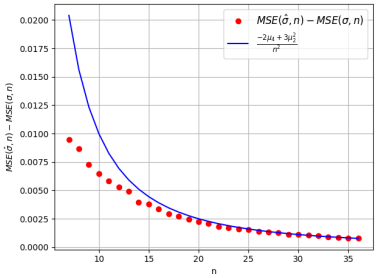
Figure 1: Linear Regression (Top-Left and Bottom-Left) and Regression Tree (Top-Right and Bottom-Right) Predictors. We observe the $O(\frac{1}{N})$ convergence rate that we demonstrated analytically. We also observe that for the Linear Regression, considered as a stable predictor, bagging does not help. On the other hand, for the Decision Tree Regression, considered as an unstable predictor (following the definition of (Breiman, 1996), predictors are unstable if a small change in the training set can result in large changes in predictions), bagging helps.



Distance Between MSE of a bagged and non-bagged estimator for a standard Gaussian distribution, averaged over 10000. Here $-2\mu_4 + 3\mu_2^2 = -3$.



Distance Between MSE of a bagged and non-bagged estimator for a Uniform disribution between -1 and 1, averaged over 10000. Here $-2\mu_4 + 3\mu_2^2 = -\frac{1}{15}$.



Distance Between MSE of a bagged and non-bagged estimator for a Rademacher disribution, averaged over 10000. Here $-2\mu_4 + 3\mu_2^2 = 1$.

Figure 2: Usual Distribution Experiments

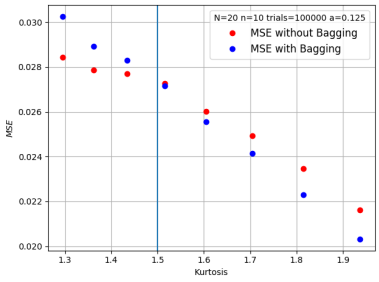


Figure 3: The MSE of the variance estimator of a distribution with or without bagging with respect to the kurtosis. As expected, we observe that the MSE of the bagged variance estimator becomes lower than the MSE of the non-bagged variance estimator after $\frac{3}{2}$

References

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

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

Andreas Buja and Werner Stuetzle. Observations on bagging. *Statistica Sinica*, pages 323–351, 2006.

Andreas Buja and Werner Stuetzle. Smoothing effects of bagging: Von mises expansions of bagged statistical functionals. *arXiv preprint arXiv:1612.02528*, 2016.

Eungchun Cho and Moon Jung Cho. Variance of sample variance with replacement. *International Journal of Pure and Applied Mathematics*, 52(1):43–47, 2009.

Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

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

Alireza Fazelpour, Taghi M Khoshgoftaar, David J Dittman, and Amri Naplitano. Investigating the variation of ensemble size on bagging-based classifier performance in imbalanced bioinformatics datasets. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pages 377–383. IEEE, 2016.

Jerome H Friedman and Peter Hall. On bagging and nonlinear estimation. *Journal of statistical planning and inference*, 137(3):669–683, 2007.

Yves Grandvalet. Bagging equalizes influence. *Machine Learning*, 55(3):251–270, 2004.

Eric Jones, Travis Oliphant, and Pearu Peterson. Scipy: Open source scientific tools for python. 2001.

Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

Aurélie Lemmens and Christophe Croux. Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2):276–286, 2006.

Luoluo Liu, Trac D Tran, et al. Reducing sampling ratios and increasing number of estimates improve bagging in sparse regression. *arXiv preprint arXiv:1812.08808*, 2018.

Jianchang Mao. A case study on bagging, boosting and basic ensembles of neural networks for ocr. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, volume 3, pages 1828–1833. IEEE, 1998.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Colin Rose and Murray D Smith. Mathstatica: mathematical statistics with mathematica. In *Compstat*, pages 437–442. Springer, 2002.

Marina Skurichina and Robert PW Duin. Bagging for linear classifiers. *Pattern Recognition*, 31(7):909–930, 1998.

Daria Sorokina, Rich Caruana, and Mirek Riedewald. Additive groves of regression trees. In *European Conference on Machine Learning*, pages 323–334. Springer, 2007.

Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

Geoffrey I Webb and Zijian Zheng. Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. *IEEE Transactions on Knowledge and Data Engineering*, 16(8):980–991, 2004.

Song Xi Chen and Peter Hall. Effects of bagging and bias correction on estimators defined by estimating equations. *Statistica Sinica*, pages 97–109, 2003.