The Effect of Bagging on Variance, Bias, and Mean Squared Error

Andreas Buja[1] Werner Stuetzle[2]

October 19, 2000

Abstract

Bagging is a device intended for reducing the prediction error of learning algorithms. Bagging consists of drawing bootstrap samples from the training sample, applying the learning algorithm to each bootstrap sample, and averaging the resulting prediction rules. Heuristically, the averaging process should reduce the variance component of the prediction error. We study the effects of bagging for U-statistics of any order and finite sums thereof. U-statistics of high order can describe complex data dependences and yet they admit a rigorous asymptotic analysis. The following are some findings:

- The effects of bagging on variance, squared bias and mean squared error are of order N^{-2}. (The following statements are all meant to second order.)
- If one allows boostrap samples *with or without* replacement and arbitrary resample sizes, then bagging based on "sampling with" for resample size M_{with} is equivalent to "sampling without" for resample size $M_{w/o}$ if $N/M_{with} = N/M_{w/o} - 1 \to g\ (> 0,\ < \infty)$.
- Var(bagged)−Var(raw) is a linear function of g; bagging improves variance if the slope is negative.
- Bias ^{2}(bagged) − Bias ^{2}(raw) is a positive quadratic function of g; bagging hence always increases squared bias.
- MSE ^{2}(bagged) − MSE ^{2}(raw) is a quadratic function of g; bagging may or may not improve mean squared error, and if it does, it is for sufficiently small g, that is, sufficiently large resample sizes M.

[1] AT&T Labs–Research, 180 Park Ave, Florham Park, NJ 07932-0971; andreas@research.att.com

[2] Department of Statistics, University of Washington, Seattle, WA 98195-4322; wxs@stat.washington.edu. Research partially supported by NSF grant DMS - 9803226. This work was performed while the second author was on sabbatical leave at AT&T Labs.

1 Introduction

Bagging, short for "bootstrap aggregation", was introduced by Breiman (1996) as a device for reducing the prediction error of learning algorithms. Bagging is performed by drawing bootstrap samples from the training sample, applying the learning algorithm to each bootstrap sample, and averaging/aggregating the resulting prediction rules, that is, averaging or otherwise aggregating the predicted values for test observations. Breiman presents empirical evidence that bagging can indeed reduce prediction error. It appears to be most effective for CART trees (Breiman *et al.* 1984). Breiman's heuristic explanation is that CART trees are highly unstable functions of the data — a small change in the training sample can result in a very different tree — and that averaging over bootstrap samples reduces the variance component of the prediction error.

In a recent report, Friedman and Hall (2000) present an asymptotic analysis of bagging purporting to explain its effects for very general statistics. The generality of their results comes at the cost of reduced transparency. One purpose of this note is to understand what Friedman and Hall could have meant.

We study the effects of bagging for the class of U-statistics of any order and finite sums thereof. While these do not capture the statistical properties of CART trees (see Buhlmann and Yu (2000) for a more realistic approach), U-statistics can capture complex interactions and yet a rigorous second order analysis is possible.

Like Friedman and Hall's, our analysis covers several variations on bagging. Instead of averaging the values of a statistic over bootstrap samples of the same size N as the original sample, we may choose the resample size M to be smaller, or even larger, than N. Another alternative covered by our analysis is resampling without replacement. We show that there exists a correspondence between resample sizes for sampling with and without replacement that renders the two sampling modes second order equivalent. We give theoretical conditions under which bagging improves the variance or the MSE of sums of U-statistics. Furthermore, under these conditions we give the range of resample sizes for which bagging yields improvements.

The correspondence between resample sizes for sampling with and without replacement for the variance is present in Section 2.6 of Friedman and Hall (2000), but it is not made use of except for the second order equivalence of conventional bootstrap ($M_{with} = N$) and half-sampling ($M_{w/o} = N/2$).

2 Resampling for U-Statistics

Let $X_1, X_2, \ldots, X_N$ be i.i.d. random variables. We consider statistics of $X_1, ..., X_N$ that are finite sums

$$U \;=\; \frac{1}{N} \sum_i A_{X_i} \;+\; \frac{1}{N^2} \sum_{i,j} B_{X_i,X_j} \;+\; \frac{1}{N^3} \sum_{i,j,k} C_{X_i,X_j,X_k} \;+\; \ldots$$

of U-statistics B, C,... that are permutation symmetric in their arguments. [We put the arguments in subscripts in order to avoid the clutter caused by frequent parentheses.] The normalizations of the sums are such that under common assumptions limits for $N \to \infty$ exist. Strictly speaking, only the off-diagonal part $\sum_{imj} B_{X_i,X_j}$ (e.,g.) is a proper U-statistic. Because we include the diagonal $i = j$ in the double sum, this is strictly speaking a V-statistic (Serfling 1980), but we use the better known term "U-statistic" anyway. The reason for including the diagonal is that only in this way can U be interpreted as the plug-in estimate $U(F_N)$ of a statistical functional

$$U(F) \;=\; \mathbf{E}\, A_X \;+\; \mathbf{E}\, B_{X,Y} \;+\; \mathbf{E}\, C_{X,Y,Z} \;+\; \ldots$$

where X, Y, Z,... are i.i.d. Knowing what the statistic U estimates is a necessity for bias calculations. A second reason for including the diagonal is that bagging has the effect of introducing terms such as B_{X_i,X_i}, so we may as well include such terms from the outset.

It is possible to explicitly calculate the bagged version of the statistic U, denoted U^{bag}. To this end, let $\mathbf{W} = W_1 \ldots, W_N \geq 0$ be integer valued random variables counting the multiplicities of $X_1, \ldots, X_N$ in a *resample*. We allow both resampling *with* and *without* replacement, and we also allow the resample size M to be arbitrary:

- For resampling *with* replacement, that is, *bootstrap*, the distribution of $\mathbf{W}$ is Multinomial$(1/N, \ldots, 1/N; M)$. Conventional bootstrap is for $M = N$, but we allow M to range between 1 and ∞. Although $M > N$ is computationally undesirable, infinity is the conceptually plausible upper bound on M: for $M = \infty$ no resampling takes place because with an "infinite resample" one knows the resampling distribution.

- For sampling *without* replacement, that is, *subsampling*, the distribution of $\mathbf{W}$ is Hypergeometric(M, N). Half-sampling, for example, is for $M = N/2$, but the resample size M can range between 1 and N. For the upper bound $M = N$ no resampling takes place because the subsample is just a permutation of the data.

With these facts we can write down the resampled and the bagged version of a U-statistic explicitly. We illustrate this for statistics with terms A_{X_i} and B_{X_i,X_j}:

$$U^{boot} \;=\; \frac{1}{M} \sum W_i \, A_{X_i} \;+\; \frac{1}{M^2} \sum W_i W_j \, B_{X_i,X_j} \;.$$

The bagged version of U under either mode of resampling is the expected value with respect to $\mathbf{W}$:

$$\begin{aligned} U^{bag} &= \mathbf{E}_{\mathbf{W}} \left[\frac{1}{M} \sum W_i \; A_{X_i} \; + \; \frac{1}{M^2} \sum W_i W_j \; B_{X_i,X_j} \right] \\ &= \frac{1}{M} \sum \mathbf{E}\left[W_i\right] \, A_{X_i} \; + \; \frac{1}{M^2} \sum \mathbf{E}\left[W_i W_j\right] \, B_{X_i,X_j} \; . \end{aligned}$$

From the form of U^{bag} it is apparent that the only relevant quantities are moments of $\mathbf{W}$:

$$\begin{aligned} \mathbf{E}\, W_i &= \frac{M}{N} & & \text{with and w/o} \\ \mathbf{E}\, W_i^2 &= \begin{cases} \text{with:} & \frac{M}{N}\left(1 + \frac{M-1}{N}\right) \\ \text{w/o:} & \frac{M}{N} \end{cases} \\ \mathbf{E}\, W_i W_j &= \begin{cases} \text{with:} & \frac{M(M-1)}{N^2} \\ \text{w/o:} & \frac{M(M-1)}{N(N-1)} \end{cases} & & (i \neq j) \end{aligned}$$

The required moments are of the same order as the order of the U-statistic. The bagged functional can now be written down explicitly. It is necessary to distinguish between the two resampling modes: we denote U^{bag} by U^{with} and $U^{w/o}$ for resampling with and without replacement, respectively.

$$\begin{aligned} U^{with} &= \frac{1}{N} \sum_i \left(A_{X_i} + \frac{1}{M} B_{X_i,X_i} \right) \; + \; \frac{1}{N^2} \sum_{i,j} \left(1 - \frac{1}{M} \right) B_{X_i,X_j} \; , \\ U^{w/o} &= \frac{1}{N} \sum_i \left(A_{X_i} + \left(\frac{1 - \frac{M}{N}}{1 - \frac{1}{N}} \right) \frac{1}{M} \, B_{X_i,X_i} \right) \; + \; \frac{1}{N^2} \sum_{i,j} \left(\frac{1 - \frac{1}{M}}{1 - \frac{1}{N}} \right) B_{X_i,X_j} \; . \end{aligned}$$

Such calculations can be extended to statustics with U-terms of higher order than two. We summarize:

Proposition 1: *A bagged sum of U-statistics is also a sum of U-statistics. For a statistic with terms* A *and* B *only, the bagged terms* A_x^{with}, $B_{x,y}^{with}$ *and* $A_x^{w/o}$, $B_{x,y}^{w/o}$, *respectively, depend on* A_x *and* $B_{x,y}$ *as follows:*

$$\begin{aligned} A_x^{with} &= A_x + \frac{1}{M} B_{x,x} \; , & B_{x,y}^{with} &= \left(1 - \frac{1}{M} \right) B_{x,y} \; , \\ A_x^{w/o} &= A_x + \left(\frac{1 - \frac{M}{N}}{1 - \frac{1}{N}} \right) \frac{1}{M} \, B_{x,x} \; , & B_{x,y}^{w/o} &= \left(\frac{1 - \frac{1}{M}}{1 - \frac{1}{N}} \right) B_{x,y} \; . \end{aligned}$$

We see that the effect of bagging is to remove mass from the proper U-part of B ($\sum_{i \neq j}$) and shift it to the diagonal ($\sum_{j=k}$), thus increasing the importance of the additive part. Similar effects take place in higher orders where variability is shifted to lower orders.

Notation: For U-statistics $C_{X,Y,Z,\ldots}$ of any order we denote partial conditional expectations by

$$C_X = \mathbf{E}\,[\,C_{X,Y,Z,W,\ldots}|\,X\,]\;,\quad C_{X,Y} = \mathbf{E}\,[\,C_{X,Y,Z,W,\ldots}|\,X,Y\,]\;,\quad C_{X,Y,Z} = \mathbf{E}\,[\,C_{X,Y,Z,W,\ldots}|\,X,Y,Z\,]\;.$$

Equivalently one could introduce them as partial marginal expectations:

$$C_x = \mathbf{E}\,[\,C_{x,Y,Z,W,\ldots}]\;,\quad C_{x,y} = \mathbf{E}\,[\,C_{x,y,Z,W,\ldots}]\;,\quad C_{x,y,z} = \mathbf{E}\,[\,C_{x,y,z,W,\ldots}]\;.$$

It will turn out that for variance and bias calculations to order N^{-2} these three partial conditional expectations are the only information needed about a U-statistic of any order. We will use simple facts such as the following without further mention:

$$\begin{aligned}\mathrm{Cov}(B_{X,Y,Z,\ldots}, C_{X,Y',Z',\ldots}) &= \mathrm{Cov}(B_X, C_X)\;,\\ \mathrm{Cov}(C_{X,Y,Z,\ldots}, C_{X,Y',Z',\ldots}) &= \mathrm{Var}(C_X)\;.\end{aligned}$$

3 Variance

Variances of finite sums of U-statistics can be calculated explicitly. For example, for a statistic that has only terms A and B, the variance is

$$\begin{aligned}\mathrm{Var}(U) &= N^{-1}\,\mathrm{Var}(A_X + 2B_X)\\ &+ N^{-2}\;(2\mathrm{Cov}(A_X, B_{X,X}) + 4\mathrm{Cov}(B_{X,X}, B_X) - 4\mathrm{Cov}(A_X, B_X)\\ &\qquad +2\mathrm{Var}(B_{X,Y}) - 12\mathrm{Var}(B_X))\\ &+ N^{-3}\;(\mathrm{Var}(B_{X,X}) - 2\mathrm{Var}(B_{X,Y}) + 8\mathrm{Var}(B_X) - 4\mathrm{Cov}(B_{X,X}, B_X))\end{aligned}$$

What matters here, though, are not variances, but differences between variances of bagged and raw statistics:

Proposition 2: *Let $g = \frac{N}{M}$ for sampling with replacement and $g = \frac{N}{M} - 1$ for sampling without replacement. Assume that these quantities stay bounded away from zero and infinity as $N \to \infty$. Let U be a finite sum of U-statistics, then:*

$$\mathrm{Var}(U^{bag}) - \mathrm{Var}(U) \;=\; \frac{1}{N^2}\cdot 2\,T_{\mathrm{Var}}\cdot g \;+\; O(\frac{1}{N^3})\;.$$

for both sampling with and without replacement. If U has only terms A_X and $B_{X,Y}$, then:

$$T_{\mathrm{Var}} \;=\; \mathrm{Cov}(A_X + 2B_X, B_{X,X} - B_X)\;.$$

The proof is in the appendix, sections A1-A4. Section A4 shows how to calculate T_{Var} for statistics with U-terms of any order.

The assumption about g is essential. If it is not satisfied, the order of terms in powers of N^{-1} will be affected. The jackknife is a case in point: it is obtained for $M = N-1$ and sampling without replacement, which does not satisfy the assumption of the proposition. It would be easy to cover such cases because the calculations we performed can be performed exact although we report them only to N^{-2}.

In bagged and unbagged functionals, the terms of order $\frac{1}{N}$ are identical and cancel out (Friedman and Hall 2000), hence bagging has no effect on the variance to order $\frac{1}{N}$. Surprisingly, the terms of order N^{-2} are identical for sampling with and without replacement modulo differing interpretations of g, a fact that will be observed again for squared bias and hence MSE as well.

There exist situations in which bagging is detrimental for the variance: $T_{\mathrm{Var}} > 0$. Bagging reduces variance iff $T_{\mathrm{Var}} < 0$. Under this condition, the variance is reduced the more the larger g and hence the smaller the resample size M is. Therefore, the fact that bagging may reduce variance cannot be the whole story: if variance were the criterion of interest, one should choose the resample size M always as low as possible for maximal variance reduction. Obviously, one has to take into account bias as well.

4 Bias

The result of this section is that bagging *always* increases squared bias for sums of U-statistics. Recall that the statistic $U = U(F_N)$ is the plug-in estimator for the functional $U(F)$, so the bias is $\mathbf{E}\; U(F_N) - U(F)$.

Proposition 3: *With the same assumptions as in Proposition 2, we have:*

$$\mathrm{Bias}^2(U^{bag}) - \mathrm{Bias}^2(U) \;=\; \frac{1}{N^2}\,(g^2+2g)\,T_{\mathrm{Bias}} \;+\; O(\frac{1}{N^3}) \;,$$

for both sampling with and without replacement. If U has only terms A_X and $B_{X,Y}$, then

$$T_{\mathrm{Bias}} \;=\; \left(\mathbf{E}\,B_{X,X} - \mathbf{E}\,B_{X,Y}\right)^2 \;.$$

The appendix, section A5, has proofs and a general formula for T_{Bias} for statistics with U-terms of any order.

Just as in the comparison of variances, sampling with and without replacement agree in the N^{-2} term modulo differing interpretation of g in the two resampling modes.

5 Mean Squared Error

The mean squared error of $U = U(F_N)$ is

$$MSE(U) \;=\; \mathbf{E}\left([U(F_N) - U(F)]^2\right) \;=\; \mathrm{Var}(U) + \mathrm{Bias}\,(U)^2\ .$$

The difference between MSEs of bagged and raw functionals is as follows:

Proposition 4: *With the same notations and assumptions as in Propositions 2 and 3, we have:*

$$MSE(U_M^{bag}(F_N)) - MSE(U(F_N)) \;=\; \frac{1}{N^2}\left(T_{\text{Bias}}\, g^2 + (T_{\text{Var}} + T_{\text{Bias}})\, 2g\right) \;+\; O(\frac{1}{N^3})\ .$$

for both sampling with and without replacement.

6 Comparison of Sampling With and Without Replacement

Variance, squared bias and hence MSE of bagged sums of U-statistics all agree in the N^{-2} term under the correspondence $g_{with} = g_{w/o}$, where $g_{with} = N/M_{with}$ and $g_{w/o} = N/M_{w/o} - 1$. This correspondence is more intuitive if one expresses the resample sizes M_{with} and $M_{w/o}$ as fractions of the sample size N:

$$\alpha_{with} = \frac{M_{with}}{N} \quad (>0,\ \ <\infty) \quad \text{and} \qquad \alpha_{w/o} = \frac{M_{w/o}}{N} \quad (>0,\ \ <1).$$

The condition $g_{with} = g_{w/o}$ is equivalent to

$$\alpha_{with} = \frac{\alpha_{w/o}}{1 - \alpha_{w/o}}\ .$$

It equates, for example, half-sampling w/o replacement, $\alpha_{w/o} = 1/2$, with conventional bootstrap, $\alpha_{with} = 1$. Subsampling w/o replacement would be natural with $\alpha_{w/o} > 1/2$ also, but it corresponds to bootstrap with $\alpha_{with} > 1$, i.e., bootstrap samples larger than the data sample. While this may be computationally not viable, it is natural to allow for this possibility if only to complete the range for bootstrap samples corresponding to subsample sizes $M_{w/o} > N/2$. The correspondence maps $\alpha_{w/o} = 1$ to $\alpha_{with} = \infty$, both of which mean that no resampling takes place.

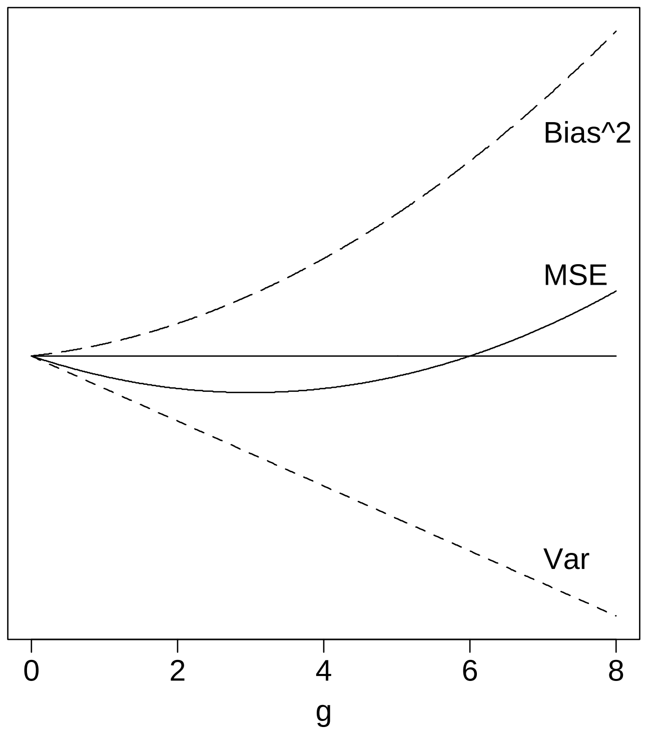

Figure 1: Dependence of Variance, Squared Bias and MSE on g. The graph shows the situation for $T_{\text{Var}}/T_{\text{Bias}} = -4$. Bagging is beneficial for $g < 6$, that is, for resample sizes $M_{with} > N/6$ and $M_{w/o} > N/7$. Optimal is $g = 3$, that is, $M_{with} = N/3$ and $M_{w/o} = N/4$.

7 Choice of Resample Size

In some situations one may obtain a gain in MSE for some resample sizes M but not for others, while in other situations one obtains no gain for any resample size. Critical is the dependence of the N^{-2} term of the MSE difference on g:

$$T_{\text{Bias}}\, g^2 \;+\; 2\,(T_{\text{Var}} + T_{\text{Bias}})\, g \;.$$

One immediately reads off the following condition for MSE improvement:

Corollary 5: *There exist resample sizes for which bagging improves the MSE to order N^{-2} iff*

$$T_{\text{Var}} + T_{\text{Bias}} < 0 \;.$$

Under this condition the range of beneficial resample sizes is characterized by

$$g < -2\left(\frac{T_{\text{Var}}}{T_{\text{Bias}}} + 1\right) \ .$$

The resample size with optimal MSE improvement is

$$g^{opt} = -\left(\frac{T_{\text{Var}}}{T_{\text{Bias}}} + 1\right) \ .$$

Conventional bootstrap, $M_{with} = N$, and half-sampling, $M_{w/o} = N/2$, (both characterized by $g = 1$) are beneficial iff

$$\frac{T_{\text{Var}}}{T_{\text{Bias}}} < -\frac{3}{2} \ ,$$

and they are optimal iff

$$\frac{T_{\text{Var}}}{T_{\text{Bias}}} = -2 \ .$$

Recall from Proposition 2 that the resample sizes M_{with} and $M_{w/o}$ are expressed in terms of $g_{with} = N/M_{with}$ and $g_{w/o} = N/M_{w/o} - 1$. The corollary therefore prescribes a minimum resample size in order to achieve MSE improvement. See Figure 1 for an illustration.

The intuition that the benefits of bagging arise from variance reduction is thus correct, although it must be qualified: Bagging is not always beneficial, but if it is, the reduction in MSE is due to reduction in variance. This follows from the fact that T_{Bias} is always positive, hence bagging always increases bias, but if the variance dips sufficiently strongly, an overall benefit results.

Recall that strictly speaking the above statements should be limited to areas bounded away from zero and infinity. Near either boundary a different type of asymptotics sets in.

8 An Example

Consider as a concrete example of U-statistics the case of quadratic functions: $A_X = a \cdot X^2$ and $B_{X,Y} = b \cdot XY$, i.e.,

$$U = a \cdot \frac{1}{N}\sum X_i^2 + b \cdot (\frac{1}{N}\sum X_i)^2 \ .$$

In order to determine the terms T_{Var} and T_{Bias}, we need the first four moments of X: Let $\mu = \mathbf{E}\,X$, $\sigma^2 = \mathbf{E}\,[(X-\mu)^2]$, $\gamma = \mathbf{E}\,[(X-\mu)^3)]/\sigma^3$ and $\kappa = \mathbf{E}\,[(X-\mu)^4]/\sigma^4$ be expectation, variance, skewness and kurtosis, respectively. Then:

$$T_{\text{Var}} \ = \ (2\mu\gamma\sigma^3 \ + \ (\kappa - 1)\sigma^4)\,ab \ + \ 2\mu\gamma\sigma^3\,b^2$$

and

$$T_{\text{Bias}} \;=\; b^2\,\sigma^4 \;.$$

It is convenient to write the criterion for the existence of resample sizes with beneficial effect as $T_{\text{Var}}/T_{\text{Bias}} + 1 < 0$:

$$\left(2\frac{\mu}{\sigma}\gamma \;+\; (\kappa-1)\right)\,\frac{a}{b} \;+\; \left(2\frac{\mu}{\sigma}\gamma \;+\; 1\right) \;<\; 0\;.$$

If $\mu = 0$ or $\gamma = 0$, this simplifies to

$$(\kappa-1)\,\frac{a}{b} \;+\; 1 \;<\; 0\;.$$

Since $\kappa > 1$ for all distributions except a balanced 2-point mass, the condition becomes

$$\frac{a}{b} \;<\; -\frac{1}{\kappa-1}\;.$$

For $a = 1,\ b = -1$, i.e., the empirical variance $U = \text{mean}(X^2) - \text{mean}(X)^2$, beneficial effects of bagging exist iff $\kappa > 2$. For $a = 0$, i.e., the squared mean $U = \text{mean}(X)^2$, no beneficial effects exist.

9 Conclusions

The major factual conclusions are stated in the abstract and do not need to be repeated here. There remains the question of what the facts mean.

Friedman and Hall (2000) seem to imply that a more general conclusion can be drawn for variance reduction. They argue that in situations with many parameters the typical effect is that bagging reduces the variance of higher order terms of a polynomial approximation to the statistic of interest; they imply that this reduction is by a full order to N^{-3}. Based on the calculations above, we are unable to confirm these conclusions. According to Proposition 1 above, the reduction in the second order U-term is minimal: from B to $B^{bag} = \left(1 - \frac{1}{M}\right)B$. The same holds for higher order U-terms. Instead, we see the reduction in variance, if any, arise from an interplay between various terms, as in the quantity

$$T_{\text{Var}} \;=\; \text{Cov}(A_X + 2B_X, B_{X,X} - B_X)\;,$$

whose sign ($T_{\text{Var}} < 0$) determines the presence of variance reduction for $U = \sum A_{X_i} + \sum B_{X_i,X_j}$. Even when variance is reduced, it is not by a change in the order of the variance of anything.

Variance cannot be the only criterion of performance: if it were, one would choose the resample size for bagging extremely small. Variance, however, is counterbalanced by bias (squared), which is always increased by bagging. Any beneficial effect of bagging on variance has to first make up for the detrimental effect on bias. The detriment becomes the greater the smaller the resample size is. Because squared bias increases quadratically in g and variance decreases linearly in g (if at all), bias detriment will overcome any variance benefit for small resample sizes. Thus, if an overall improvement in MSE takes place, it is for resample sizes M whose g is sufficiently close to zero, that is, M_{with} sufficiently close to ∞ or $M_{w/o}$ sufficiently close to N. The strange situation can arise where beneficial resample sizes M_{with} can be found only above a threshold that is larger than N.

References

[1] L. Breiman. Bagging predictors. *Machine Learning*, 26:123–140, 1996.

[2] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, California, 1984.

[3] Peter Buhlmann and Bin Yu. Explaining bagging. Can be downloaded from http://stat.ethz.ch/˜buhlmann/bibliog.html, September 2000.

[4] J.H. Friedman and O. Hall. On bagging and nonlinear estimation. Can be downloaded from http://www-stat.stanford.edu/˜jhf/#reports, May 2000.

[5] R. J. Serfling. *Approximation Theorems of Mathematical Statistics.* Wiley, New York, 1980.

Appendix

We derive the results in steps:

A1. Covariance of General Interactions

We introduce notation for statistical functionals that are interactions of order J and K, respectively:

$$\mathbf{B} \;=\; \frac{1}{N^J} \sum_{\mu} B_\mu \;, \qquad \mathbf{C} \;=\; \frac{1}{N^K} \sum_{\nu} C_\nu \;,$$

where

$$\mu = (\mu_1, \ldots, \mu_J) \in \{1, \ldots, N\}^J \;, \qquad B_\mu \;=\; B_{X_{\mu_1}, \ldots, X_{\mu_J}} \;,$$
$$\nu = (\nu_1, \ldots, \nu_K) \in \{1, \ldots, N\}^K \;, \qquad C_\nu \;=\; C_{X_{\nu_1}, \ldots, X_{\nu_K}} \;.$$

We assume the functions $B_{x_1,\ldots,x_J}$ and $C_{y_1,\ldots,y_K}$ to be permutation symmetric in their arguments, the random variables $X_1, \ldots, X_N$ to be i.i.d., and the second moments of B_μ and C_ν to exist for all μ and ν. As is usual in the context of von Mises expansions, we do not limit the summations to distinct indices as is usual in the context of U-statistics. The reason is that we wish $\mathbf{B}$ and $\mathbf{C}$ to be plug-in estimates of the functionals $\mathbf{E}\, B_{1,\ldots,J}$ and $\mathbf{E}\, C_{1,\ldots,K}$.

We are interested in calculating the covariance between $\mathbf{B}$ and $\mathbf{C}$ to order N^{-1} and N^{-2}. To this end, we need some additional notation for index calculations: Let $t(\mu)$ and $t(\nu)$ be the number of essential ties in μ and ν, respectively, as follows:

$$t(\mu) \;=\; \#\{\; (i,j) \mid i < j,\; \mu_i = \mu_j \;,\;\; \mu_i \neq \mu_1, \ldots, \mu_{i-1} \;\} \;.$$

The index i marks the first appearance of the value μ_i, and all other μ_j equal to μ_i are counted relative to i. For example, $\mu = (1,1,2,1,2)$ has three essential ties: $\mu_1 = \mu_2$, $\mu_1 = \mu_4$, and $\mu_3 = \mu_5$; the tie $\mu_2 = \mu_4$ is inessential because it can be inferred from the essential ties.

Another notation we need is for the number $c(\mu,\nu)$ of essential cross-ties between μ and ν:

$$c(\mu,\nu) \;=\; \#\{\; (i,j) \mid \mu_i = \nu_j \;,\;\; \mu_i \neq \mu_1, \ldots, \mu_{i-1} \;, \nu_j \neq \nu_1, \ldots, \nu_{j-1} \;\} \;.$$

We exclude inessential cross-ties that can be inferred from the ties within μ and ν. For example, for $\mu = (1,2,1)$ and $\nu = (3,1)$ the only essential cross-tie is $\mu_1 = \nu_2 = 1$; the remaining inessential cross-tie $\mu_3 = \nu_2$ can be inferred from the essential tie $\mu_1 = \mu_3$ within μ.

With these definitions we have the following fact for the number of essential ties of the concatenated sequence (μ,ν):

$$t((\mu,\nu)) \;=\; t(\mu) + t(\nu) + c(\mu,\nu) \;.$$

In expanding the covariance between $\mathbf{B}$ and $\mathbf{C}$, we note that the terms with zero cross-ties between μ and ν vanish due to independence. Thus:

$$\mathrm{Cov}(\mathbf{B},\mathbf{C}) \;=\; \frac{1}{N^{J+K}} \sum_{c(\mu,\nu)>0} \mathrm{Cov}(B_\mu, C_\nu) \;.$$

Because $\#\{(\mu,\nu) \mid c(\mu,\nu) > 0\ \}$ is of order $O(N^{J+K-1})$ (a crude upper bound is JKN^{J+K-1}), it follows that $\mathrm{Cov}(\mathbf{B},\mathbf{C})$ is of order $O(N^{-1})$, as it should.

We now show that in order to capture terms of order N^{-1} and N^{-2} in $\mathrm{Cov}(\mathbf{B},\mathbf{C})$ it is sufficient to limit the summation to those (μ,ν) that satisfy either

- $t(\mu) = 0$, $t(\nu) = 0$ and $c(\mu,\nu) = 1$, or
- $t(\mu) = 1$, $t(\nu) = 0$ and $c(\mu,\nu) = 1$, or
- $t(\mu) = 0$, $t(\nu) = 1$ and $c(\mu,\nu) = 1$,

or $t(\mu) + t(\nu) = 0, 1$ and $c(\mu,\nu) = 1$ for short. To this end, we note that the number of terms with $t(\mu) + t(\nu) \geq 2$ and $c(\mu,\nu) \geq 1$ is of order N^{J+K-3}. This is seen from the following crude upper bound:

$$\begin{aligned}
& \#\{\ (\mu,\nu) \mid t(\mu) + t(\nu) \geq 2\ ,\ c(\mu,\nu) \geq 1\ \} \\
\leq\ & \#\{\ (\mu,\nu) \mid t((\mu,\nu)) \geq 3\ \} \\
\leq\ & \left(\binom{K+J}{4, K+J-4} + \binom{K+J}{3, 2, K+J-5} + \binom{J+K}{2, 2, 2, J+K-6} \right) \cdot N^{J+K-3} \;,
\end{aligned}$$

where the “choose” terms arise from choosing the index patterns $(1, 1, 1, 1)$, $(1, 1, 1, 2, 2)$ and $(1, 1, 2, 2, 3, 3)$ in all possible ways in a sequence (μ,ν) of length $K + J$; these three patterns are necessary and sufficient for $t((\mu,\nu)) \geq 3$. Using N^{J+K-3} instead of $N(N-1)\ldots(N-(J+K-4))$ makes this an upper bound.

With the assumption of finite second moments of B_μ and C_ν for all μ and ν, it follows that the sum of terms with $t(\mu) + t(\nu) \geq 2$ and $c(\mu,\nu) \geq 1$ is of order $O(N^{-3})$. Abbreviating

$$\begin{bmatrix} N \\ L \end{bmatrix} \;=\; \frac{N!}{L!} \;=\; N(N-1)\ldots(N-(L-1))$$

we have:

$$\begin{aligned}
& \mathrm{Cov}(\mathbf{B},\mathbf{C}) \\
=\ & \frac{1}{N^{J+K}} \sum_{t(\mu)+t(\nu)=0,1;\ c(\mu,\nu)=1} \mathrm{Cov}(B_\mu, C_\nu) \;+\; O(N^{-3}) \\
=\ & \frac{1}{N^{J+K}} \sum_{t(\mu)=0,\ t(\nu)=0,\ c(\mu,\nu)=1} \mathrm{Cov}(B_\mu, C_\nu)
\end{aligned}$$

$$
\begin{aligned}
&+\ \frac{1}{N^{J+K}} \sum_{t(\mu)=1,\ t(\nu)=0,\ c(\mu,\nu)=1} \mathrm{Cov}(B_\mu, C_\nu) \\
&+\ \frac{1}{N^{J+K}} \sum_{t(\mu)=0,\ t(\nu)=1,\ c(\mu,\nu)=1} \mathrm{Cov}(B_\mu, C_\nu) \\
&+\ O(N^{-3}) \\
&=\ \frac{1}{N^{J+K}}\ JK \begin{bmatrix} N \\ J+K-1 \end{bmatrix} \cdot \mathrm{Cov}(B_{(1,\ldots)}, C_{(1,\ldots)}) \\
&+\ \frac{1}{N^{J+K}}\ \binom{J}{2} KN\ \begin{bmatrix} N \\ J+K-3 \end{bmatrix} \cdot \Big(\mathrm{Cov}(B_{(1,1,\ldots)}, C_{(1,\ldots)}) + \mathrm{Cov}(B_{(1,1,2,\ldots)}, C_{(2,\ldots)})\Big) \\
&+\ \frac{1}{N^{J+K}}\ J \binom{K}{2} N\ \begin{bmatrix} N \\ J+K-3 \end{bmatrix} \cdot \Big(\mathrm{Cov}(B_{(1,\ldots)}, C_{(1,1,\ldots)}) + \mathrm{Cov}(B_{(2,\ldots,J)}, C_{(1,1,2,\ldots)})\Big) \\
&+\ O(N^{-3})\ ,
\end{aligned}
$$

where "…" inside a covariance stands for as many *distinct other* indices as necessary. Using

$$
\begin{bmatrix} N \\ L \end{bmatrix}\ =\ N^L - \binom{L}{2} N^{L-1} + O(N^{L-2})
$$

we obtain

$$
\begin{aligned}
&\mathrm{Cov}(\mathbf{B}, \mathbf{C}) \\
&=\ \left(N^{-1} - \binom{J+K-1}{2} N^{-2} + O(N^{-3})\right)\ JK \cdot \mathrm{Cov}(B_{(1,\ldots)}, C_{(1,\ldots)}) \\
&+\ \Big(N^{-2} + O(N^{-3})\Big)\ \binom{J}{2} K \cdot \Big(\mathrm{Cov}(B_{(1,1,\ldots)}, C_{(1,\ldots)}) + \mathrm{Cov}(B_{(1,1,2,\ldots)}, C_{(2,\ldots)})\Big) \\
&+\ \Big(N^{-2} + O(N^{-3})\Big)\ J \binom{K}{2} \cdot \Big(\mathrm{Cov}(B_{(1,\ldots)}, C_{(1,1,\ldots)}) + \mathrm{Cov}(B_{(2,\ldots)}, C_{(1,1,2\ldots)})\Big) \\
&+\ O(N^{-3})\ .
\end{aligned}
$$

Collecting terms $O(N^{-3})$, the above can be written in a more sightly manner as

$$
\begin{aligned}
&\mathrm{Cov}(\mathbf{B}, \mathbf{C}) \\
&=\ \left(N^{-1} - \binom{J+K-1}{2} N^{-2}\right)\ JK \cdot \mathrm{Cov}(B_X, C_X) \\
&+\ N^{-2}\ \binom{J}{2} K \cdot (\mathrm{Cov}(B_{X,X}, C_X) + \mathrm{Cov}(B_{X,X,Y}, C_Y)) \\
&+\ N^{-2}\ J \binom{K}{2} \cdot (\mathrm{Cov}(B_X, C_{X,X}) + \mathrm{Cov}(B_X, C_{X,Y,Y})) \\
&+\ O(N^{-3}) \\
&\\
&=\ a \cdot N^{-1} + b \cdot N^{-2} + O(N^{-3})\ .
\end{aligned}
$$

A2. Moments of Resampling Coefficients

We consider sampling in terms of M draws from N objects $\{1, \ldots, N\}$ with and without replacement. The draws are M exchangeable random variables $S_1, \ldots, S_M$, where $S_i \in \{1, \ldots, N\}$. Each draw is equally likely: $P[S_i = n] = N^{-1}$, but for sampling with replacement the draws are independent; for sampling w/o replacement they are dependent and the joint probabilities are $P[S_1 = n_1, S_2 = n_2, \ldots, S_J = n_J] = \begin{bmatrix} M \\ J \end{bmatrix} / \begin{bmatrix} N \\ J \end{bmatrix}$ for distinct n_i's, and $= 0$ if ties exist among the n_i's.

For resampling one is interested in the count variables

$$W_{n,M,N} \;=\; W_n \;=\; \sum_{\mu=1,\ldots,M} 1_{[S_\mu = n]} \;,$$

where we drop M and N from the subscripts if they are fixed. We let $\mathbf{W} = \mathbf{W}_{M,N} = (W_1, \ldots, W_N)$ and recall:

- For resampling *with* replacement: $\mathbf{W} \sim \text{Multinomial}(1/N, \ldots, 1/N; M)$.
- For resampling *w/o* replacement: $\mathbf{W} \sim \text{Hypergeometric}(M, N)$.

For bagging one needs the moments of $\mathbf{W}$. Because of exchangeability of $\mathbf{W}$ for fixed M and N, it is sufficient to consider moments of the form

$$\mathbf{E}\,[W_{n=1,M,N}^{i_1}\; W_{n=2,M,N}^{i_2} \cdots W_{n=I,M,N}^{i_L}] \;.$$

The bagged interactions of the previous Subsection A1 are:

$$\mathbf{B}^{bag} \;=\; \frac{1}{M^J} \sum_{\mu_1,\ldots,\mu_k=1}^{N} \mathbf{E}\,[W_{\mu_1} \cdots W_{\mu_J}]\; B_\mu \;, \tag{1}$$

$$\mathbf{C}^{bag} \;=\; \frac{1}{M^K} \sum_{\nu_1,\ldots,\nu_k=1}^{N} \mathbf{E}\,[W_{\nu_1} \cdots W_{\nu_K}]\; C_\mu \;. \tag{2}$$

The following recursion formulae hold for $i_l \geq 1$:

$$\mathbf{E}\,[W_{n=1,M,N}^{i_1}\; W_{n=2,M,N}^{i_2} \cdots W_{n=I,M,N}^{i_L}]$$

$$= \begin{cases} \text{with}: & \frac{M}{N}\; \mathbf{E}\,[(W_{n=1,M-1,N} + 1)^{i_1 - 1}\; W_{n=2,M-1,N}^{i_2} \cdots W_{n=L,M-1,N}^{i_L}] \;, \\[2ex] \text{w/o}: & \frac{M}{N}\; \mathbf{E}\,[W_{n=2,M-1,N-1}^{i_2} \cdots W_{n=L,M-1,N-1}^{i_L}] \;. \end{cases}$$

From these one obtains the following results, all of which will be used below ($\alpha = M/N$):

$$\mathbf{E}\left[W_1^{i_1}\ W_2^{i_2}\cdots W_I^{i_L}\right] \ = \ O(1)$$

$$\mathbf{E}\left[W_1\ W_2\cdots W_L\right]$$

$$= \left\{ \begin{array}{ll} \text{with :} & \begin{bmatrix} M \\ L \end{bmatrix} / N^L \\ \\ \text{w/o :} & \begin{bmatrix} M \\ L \end{bmatrix} / \begin{bmatrix} N \\ L \end{bmatrix} \end{array} \right.$$

$$= \left\{ \begin{array}{ll} \text{with :} & \alpha^L \ - \ \alpha^L \binom{L}{2} \frac{1}{\alpha} N^{-1} \ + \ O(N^{-2}) \\ \\ \text{w/o :} & \alpha^L \ - \ \alpha^L \binom{L}{2} \left(\frac{1}{\alpha} - 1\right) N^{-1} \ + \ O(N^{-2}) \end{array} \right.$$

$$= \ \alpha^L \left(1 \ - \ \binom{L}{2}\ g\ N^{-1}\right) \ + \ O(N^{-2})$$

$$\mathbf{E}\left[W_1^2\ W_2\cdots W_{L-1}\right]$$

$$= \left\{ \begin{array}{ll} \text{with :} & \begin{bmatrix} M \\ L \end{bmatrix} / N^L \ + \ \begin{bmatrix} M \\ L-1 \end{bmatrix} / N^{L-1} \\ \\ \text{w/o :} & \begin{bmatrix} M \\ L-1 \end{bmatrix} / \begin{bmatrix} N \\ L-1 \end{bmatrix} \end{array} \right.$$

$$= \left\{ \begin{array}{ll} \text{with :} & \alpha^L \ + \ \alpha^{L-1} \ + \ O(N^{-1}) \\ \text{w/o :} & \alpha^{L-1} \ + \ O(N^{-1}) \end{array} \right.$$

$$= \ \alpha^L\ (g+1) \ + \ O(N^{-1})\ ,$$

where as always $g = \frac{1}{\alpha}$ for sampling with, and $g = \frac{1}{\alpha} - 1$ for sampling w/o, replacement.

A3. Covariances of Bagged Interactions

With this preparations, we can approach the covariance of the M-bagged version of general interactions:

$$\begin{aligned}
\mathbf{B}^{bag} &= \mathbf{E}_{\mathbf{W}} \left[\frac{1}{M^J} \sum_{\mu} W_{\mu_1} \cdots W_{\mu_J} \cdot B_\mu \right] &= \frac{1}{M^J} \sum_{\mu} \mathbf{E} \left[W_{\mu_1} \cdots W_{\mu_J} \right] \cdot B_\mu \ , \\
\mathbf{C}^{bag} &= \mathbf{E}_{\mathbf{W}} \left[\frac{1}{M^K} \sum_{\nu} W_{\nu_1} \cdots W_{\nu_K} \cdot C_\nu \right] &= \frac{1}{M^K} \sum_{\nu} \mathbf{E} \left[W_{\nu_1} \cdots W_{\nu_K} \right] \cdot C_\nu \ .
\end{aligned}$$

Bagging differentially reweights the parts of an interaction, and the result is not a pure interaction anymore but a general U-statistic. The effect of bagging is to create lower-order interactions from higher orders.

Note two facts about the moments of $\mathbf{W}$ which act as weights: 1) They depend on the structure of the ties of the indices only; for example, $\mu = (1,1,2)$ and $\mu = (3,2,3)$ have the same weights, $\mathbf{E}\left[W_1^2 W_2\right] = \mathbf{E}\left[W_3^2 W_2\right]$ due to exchangeability. 2) The moments of $\mathbf{W}$ are of order $O(1)$ in N (Appendix A2) and hence preserve the orders $O(N^{-1}$, $O(N^{-2}$, $O(N^{-3}$ of the terms considered in Appendix A1. These considerations allow us to extend the covariance calculations of Appendix A1 from raw to bagged interactions:

$$\begin{aligned}
&\mathrm{Cov}(\mathbf{B}^{bag}, \mathbf{C}^{bag}) \\
&= \frac{1}{M^{J+K}} \sum_{t(\mu)+t(\nu)=0,1;\ c(\mu,\nu)=1} \mathbf{E}\left[W_{\mu_1} \cdots W_{\mu_J}\right] \mathbf{E}\left[W_{\nu_1} \cdots W_{\nu_K}\right] \mathrm{Cov}(B_\mu, C_\nu) \\
&+ O(N^{-3}) \\
&= \frac{1}{M^{J+K}} \sum_{t(\mu)=0,\ t(\nu)=0,\ c(\mu,\nu)=1} \mathbf{E}\left[W_{\mu_1} \cdots W_{\mu_J}\right] \mathbf{E}\left[W_{\nu_1} \cdots W_{\nu_K}\right] \mathrm{Cov}(B_\mu, C_\nu) \\
&+ \frac{1}{M^{J+K}} \sum_{t(\mu)=1,\ t(\nu)=0,\ c(\mu,\nu)=1} \mathbf{E}\left[W_{\mu_1} W_{\mu_2} \cdots W_{\mu_J}\right] \mathbf{E}\left[W_{\nu_1} \cdots W_{\nu_K}\right] \mathrm{Cov}(B_\mu, C_\nu) \\
&+ \frac{1}{M^{J+K}} \sum_{t(\mu)=0,\ t(\nu)=1,\ c(\mu,\nu)=1} \mathbf{E}\left[W_{\mu_1} W_{\mu_2} \cdots W_{\mu_J}\right] \mathbf{E}\left[W_{\nu_1} W_{\nu_2} \cdots W_{\nu_K}\right] \mathrm{Cov}(B_\mu, C_\nu) \\
&+ O(N^{-3}) \\
&= \frac{1}{N^{J+K} \alpha^{J+K}} \, JK \begin{bmatrix} N \\ J+K-1 \end{bmatrix} \\
&\cdot \mathbf{E}\left[W_1 \cdots W_J\right] \mathbf{E}\left[W_1 \cdots W_K\right] \mathrm{Cov}(B_{(1,\ldots)}, C_{(1,\ldots)})
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{N^{J+K}\alpha^{J+K}} \binom{J}{2} KN \left[\genfrac{}{}{0pt}{}{N}{J+K-3} \right] \\
&\quad \cdot \mathbf{E}\left[W_1^2 W_2 \cdots W_{J-1}\right] \mathbf{E}\left[W_1 \cdots W_K\right] \Big(\mathrm{Cov}(B_{(1,1,\ldots)}, C_{(1,\ldots)}) + \mathrm{Cov}(B_{(1,1,2,\ldots)}, C_{(2,\ldots)}) \Big) \\
&+ \frac{1}{N^{J+K}\alpha^{J+K}} J \binom{K}{2} N \left[\genfrac{}{}{0pt}{}{N}{J+K-3} \right] \\
&\quad \cdot \mathbf{E}\left[W_1 \cdots W_J\right] \mathbf{E}\left[W_1^2 W_2 \cdots W_{K-1}\right] \Big(\mathrm{Cov}(B_{(1,\ldots)}, C_{(1,1,\ldots)}) + \mathrm{Cov}(B_{(2,\ldots,J)}, C_{(1,1,2,\ldots)}) \Big) \\
&+ O(N^{-3})
\end{aligned}
$$

$$
\begin{aligned}
&= JK \left(N^{-1} - \binom{J+K-1}{2} N^{-2} \right) \left(1 - \binom{J}{2} g N^{-1} \right) \\
&\quad \cdot \left(1 - \binom{K}{2} g N^{-1} \right) \mathrm{Cov}(B_X, C_X) \\
&+ \binom{J}{2} K N^{-2} (g+1) \left(\mathrm{Cov}(B_{X,X}, C_X) + \mathrm{Cov}(B_{X,X,Y}, C_Y) \right) \\
&+ J \binom{K}{2} N^{-2} (g+1) \left(\mathrm{Cov}(B_X, C_{X,X}) + \mathrm{Cov}(B_X, C_{X,Y,Y}) \right) \quad + O(N^{-3}) \\
&= \left(N^{-1} - N^{-2} \binom{J+K-1}{2} - N^{-2} \left(\binom{J}{2} + \binom{K}{2} \right) g \right) JK \, \mathrm{Cov}(B_X, C_X) \\
&+ N^{-2} \binom{J}{2} K (g+1) \left(\mathrm{Cov}(B_{X,X}, C_X) + \mathrm{Cov}(B_{X,X,Y}, C_Y) \right) \\
&+ N^{-2} J \binom{K}{2} (g+1) \left(\mathrm{Cov}(B_X, C_{X,X}) + \mathrm{Cov}(B_X, C_{X,Y,Y}) \right) \quad + O(N^{-3})
\end{aligned}
$$

A4. Difference Between Bagged and Raw Variances

Comparing the results of the Sections A3 and A1, we get:

$$
\begin{aligned}
&\mathrm{Cov}(\mathbf{B}^{bag}, \mathbf{C}^{bag}) - \mathrm{Cov}(\mathbf{B}, \mathbf{C}) \\
&= - N^{-2} \left(\binom{J}{2} + \binom{K}{2} \right) g \, JK \, \mathrm{Cov}(B_X, C_X) \\
&\qquad + N^{-2} \binom{J}{2} K g \left(\mathrm{Cov}(B_{X,X}, C_X) + \mathrm{Cov}(B_{X,X,Y}, C_Y) \right) \\
&\qquad + N^{-2} J \binom{K}{2} g \left(\mathrm{Cov}(B_X, C_{X,X}) + \mathrm{Cov}(B_X, C_{X,Y,Y}) \right) \quad + O(N^{-3})
\end{aligned}
$$

$$
\begin{aligned}
&= \; N^{-2} \; g \; \left(- \left(\binom{J}{2} + \binom{K}{2} \right) \; JK \; \mathrm{Cov}(B_X, C_X) \right. \\
&\qquad\qquad + \binom{J}{2} K \; \left(\mathrm{Cov}(B_{X,X}, C_X) + \mathrm{Cov}(B_{X,X,Y}, C_Y) \right) \\
&\qquad\qquad \left. + J \binom{K}{2} \; \left(\mathrm{Cov}(B_X, C_{X,X}) + \mathrm{Cov}(B_X, C_{X,Y,Y}) \right) \; \right) \; + \; O(N^{-3}) \\
&= \; N^{-2} \; g \; 2 \; T_{\mathrm{Var}}(\mathbf{B}, \mathbf{C}) \; + \; O(N^{-3}) \; ,
\end{aligned}
$$

where

$$
\begin{aligned}
T_{\mathrm{Var}}(\mathbf{B}, \mathbf{C}) \; &= \; \frac{1}{2} \; \left(\binom{J}{2} K \; \mathrm{Cov}(C_X, \; B_{X,X} + B_{X,Y,Y} - JB_X) \right. \\
&\qquad \left. + \binom{K}{2} J \; \mathrm{Cov}(B_X, \; C_{X,X} + C_{X,Y,Y} - KC_X) \; \right) \; .
\end{aligned}
$$

The expression for $T_{\mathrm{Var}}(\mathbf{B}, \mathbf{C})$ remains correct for J and K as low as 1, in which case one interprets $\binom{J}{2} = 0$ and $B_{X,X} = 0$ when $J = 1$, and $B_{X,Y,Y} = 0$ when $J \leq 2$, and similar for C when K =1 or 2.

The result generalizes to arbitrary finite sums of interactions

$$
\begin{aligned}
U \; &= \; \mathbf{A} + \mathbf{B} + \mathbf{C} + \ldots \\
&= \; \frac{1}{N} \sum_i A_i \; + \; \frac{1}{N^2} \sum_{i,j} B_{i,j} \; + \; \frac{1}{N^3} \sum_{i,j,k} C_{i,j,k} \; + \; \ldots \; .
\end{aligned}
$$

Because $T_{\mathrm{Var}}(\mathbf{B}, \mathbf{C})$ is a bilinear form in its arguments, the corresponding constant $T_{\mathrm{Var}}(U)$ for sums of U-statistics can be expanded as follows:

$$
\begin{aligned}
T_{\mathrm{Var}}(U) \; &= \; T_{\mathrm{Var}}(\mathbf{A}, \mathbf{A}) \; + \; 2\, T_{\mathrm{Var}}(\mathbf{A}, \mathbf{B}) \; + \; T_{\mathrm{Var}}(\mathbf{B}, \mathbf{B}) \\
&\quad + \; 2\, T_{\mathrm{Var}}(\mathbf{A}, \mathbf{C}) \; + \; 2\, T_{\mathrm{Var}}(\mathbf{B}, \mathbf{C}) \; + \; T_{\mathrm{Var}}(\mathbf{C}, \mathbf{C}) + \ldots \; ,
\end{aligned}
$$

so that

$$
\mathrm{Var}(U^{bag}) - \mathrm{Var}(U) \; = \; N^{-2}\, g \, 2\, T_{\mathrm{Var}}(U) \; + \; O(N^{-3}) \; .
$$

For example a functional consisting of first and second order terms,

$$
U \; = \; \mathbf{A} \; + \; \mathbf{B} \; = \; \frac{1}{N} \sum_i A_i \; + \; \frac{1}{N^2} \sum_{i,j} B_{i,j} \; ,
$$

yields

$$
\begin{aligned}
T_{\mathrm{Var}}(U) \; &= \; T_{\mathrm{Var}}(\mathbf{A}, \mathbf{A}) \; + \; 2\, T_{\mathrm{Var}}(\mathbf{A}, \mathbf{B}) \; + \; T_{\mathrm{Var}}(\mathbf{B}, \mathbf{B}) \\
&= \; \mathrm{Cov}(A_X, \; B_{X,X} - 2B_X) \; + \; 2\, \mathrm{Cov}(B_X, \; B_{X,X} - 2B_X) \\
&= \; \mathrm{Cov}(A_X + 2B_X, B_{X,X} - 2B_X) \; .
\end{aligned}
$$

Note that $T_{\mathrm{Var}}(\mathbf{A}, \mathbf{A}) = 0$ because bagging leaves additive statistics unchanged.

A5. Difference between Bagged and Raw Squared Bias

We consider a single Kth order interaction first, with functional and plug-in statistic

$$\begin{aligned} U(F) &= \mathbf{E}\; C_{(1,2,\ldots,K)}\ , \\ U(F_N) &= \frac{1}{N^K} \sum_{\nu_1,\ldots,\nu_K=1}^{N} C_{(\nu_1,\ldots,\nu_K)}\ . \end{aligned}$$

[Recall that C_ν and $C_{(\nu_1,\ldots,\nu_K)}$ are short for $C_{X_{\nu_1},\ldots,X_{\nu_K}}$.] The functional $U(F)$ plays the role of the parameter to be estimated by the statistic $U = U(F_N)$, so that the notion of bias applies.

We first calculate the bias for the raw statistic U and second for the bagged statistic U^{bag}. Note that $\mathbf{E}\, C_X = \mathbf{E}\; C_{1,\ldots,K} = U(F)$.

$$\begin{aligned} \mathbf{E}\,[U(F_N)] \;&=\; \frac{1}{N^K} \sum_{\nu_1,\ldots,\nu_K} \mathbf{E}\; C_{(\nu_1,\ldots,\nu_K)} \\ &= \frac{1}{N^K} \left(\begin{bmatrix} N \\ K \end{bmatrix} \mathbf{E}\; C_{(1,\ldots,K)} \;+\; \binom{K}{2} \begin{bmatrix} N \\ K-1 \end{bmatrix} \mathbf{E}\; C_{(1,1,2,\ldots,K-1)} \;+\; O(N^{K-2}) \right) \\ &= U(F) \;+\; N^{-1} \binom{K}{2} (\,\mathbf{E}\; C_{X,X} \;-\; \mathbf{E}\; C_X\,) \;+\; O(N^{-2})\ . \end{aligned}$$

Now for the bias of the bagged statistic:

$$\begin{aligned} \mathbf{E}\; U^{bag} \;&=\; \frac{1}{M^K} \sum_{\nu_1,\ldots,\nu_k=1}^{N} \mathbf{E}\,[W_{\nu_1} \cdots W_{\nu_K}]\; \mathbf{E}\; C_{(\nu_1,\ldots,\nu_K)} \\ &= \frac{1}{N^K \alpha^K} \left(\sum_{t(\nu)=0} \;+\; \sum_{t(\nu)=1} \;+\; O(N^{K-2}) \right) \\ &= \frac{1}{N^K \alpha^K} \left(\begin{bmatrix} N \\ K \end{bmatrix} \mathbf{E}\,[W_1 \cdots W_K]\, \mathbf{E}\; C_{(1,\ldots,K)} \right. \\ &\qquad\qquad \left. +\; \binom{K}{2} \begin{bmatrix} N \\ K-1 \end{bmatrix} \mathbf{E}\,[W_1^2 W_2 \cdots W_{K-1}]\, \mathbf{E}\; C_{(1,1,2,\ldots,K-1)} \right) \\ &\quad + O(N^{-2}) \\ &= \left(1 - \binom{K}{2} N^{-1} \right) \left(1 - \binom{K}{2}\; g\; N^{-1} \right)\; \mathbf{E}\; C_{(1,\ldots,K)} \\ &\quad + N^{-1} \binom{K}{2}\; (g+1)\; \mathbf{E}\; C_{(1,1,2,\ldots,K-1)} \\ &\quad + O(N^{-2}) \\ &= \; U(F) \;-\; N^{-1} \binom{K}{2}\; (g+1)\; \mathbf{E}\; C_{(1,\ldots,K)} \end{aligned}$$

$$+ \, N^{-1} \binom{K}{2} \;\; (g+1) \; \mathbf{E} \; C_{(1,1,2,\ldots,K-1)} \; + \; O(N^{-2})$$

$$= \;\; U(F) \; + \; N^{-1} \binom{K}{2} \;\; (g+1) \;\; (\mathbf{E} \; C_{X,X} \; - \; \mathbf{E} \; C_X) \;\; + \; O(N^{-2})$$

Thus:

$$\text{Bias}\,(U^{bag}) \;\; = \;\; N^{-1} \binom{K}{2} \;\; (g+1) \;\; (\mathbf{E} \; C_{X,X} \; - \; \mathbf{E} \; C_X) \;\; + \; O(N^{-2})$$

As for variances, we can now consider more general statistics that are finite sums of interactions:

$$\begin{aligned} U \;\; &= \;\; \mathbf{A} \; + \; \mathbf{B} \; + \; \mathbf{C} \; + \; \ldots \\ b \;\; &= \;\; \frac{1}{N} \sum A_i \; + \; \frac{1}{N^2} \sum B_{i,j} \; + \; \frac{1}{N^3} \sum C_{i,j,k} \; + \; \ldots \end{aligned}$$

The final result is:

$$\begin{aligned} &\text{Bias}^{\,2}(U^{bag}) \; - \; \text{Bias}^{\,2}(U) \\ &= \; N^{-2} \;\; \Big((g+1)^2 - 1\Big) \left(\binom{2}{2} (\mathbf{E} \; B_{X,X} \; - \; \mathbf{E} \; B_X) \; + \; \binom{3}{2} (\mathbf{E} \; C_{X,X} \; - \; \mathbf{E} \; C_X) \; + \; \ldots \right)^2 \\ &\quad + \; O(N^{-3}) \; . \end{aligned}$$

As usual, $g = \frac{1}{\alpha}$ for sampling with, and $g = \frac{1}{\alpha} - 1$ for sampling w/o, replacement.