Another look at Breiman's Splitting Criteria

Qiming Du

February 25, 2020

Contents

1 Story and background

In the meeting of 20 February 2020, Gérard suggested that I look into the performance w.r.t. Wasserstein distance of conditional distribution estimation given by Breiman's original Random Forests (RF). Apriori, it does not work: the splitting criterion of Breiman's RF seems has nothing to do with Wasserstein distance. However, numeric never lies, and I was surprised that RF systematically outperform the original Wasserstein RF in almost every way. To be clearer, by original Wasserstein RF, we mean that the splitting is conducted by maximize the Wasserstein distance of empirical measures associated to A_L and A_R; and by outperform we mean that the Average Wasserstein Distance (AWD) tested on 500 points in test dataset is smaller. Therefore, it is natural to suppose that Breiman's splitting rule has some intrinsic connection with Waserstein distance. By studying it, we expect to have a deeper understanding in the following aspects:

- A natural generalization which works in a multivariate setting, i.e., the dimension of the objective Y_i is bigger than 1;
- A modification of Wasserstein RF such that it can systematically outperform (at least by strong numerical evidence) the Breiman's RF.

2 Notation

If not mentioned, the state space is $\mathbf{R}^d$ and $|\cdot|$ denotes the Euclidean norm. We consider a dataset $(X_i, Y_i; 1 \le i \le N)$. We denote A_L and A_R the left node and right node w.r.t. some split of data points $A := A_L \cup A_R$. We also denote $N_A := N$, $N_L := \#(A_L)$ and $N_R := \#(A_R)$. We denote π_L, π_R and π_A the empirical measures w.r.t. Y_i associated to A_L, A_R and A. The mechanism of Random Forests (RF) corresponds to the classical version of Breiman's original RF with axis-aligned cuts. We only discuss the modification of splitting criteria. Said differently, a splitting criteria defines a variant of Random Forest.

3 Wasserstein distance between empirical measures

A natural metric to measure the distance of two measures is Wasserstein distance. More precisely, the Wasserstein distance $\mathcal{W}_p$ between two measures μ, ν on $\mathbf{R}^d$ is defined by

$$\mathcal{W}_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int |x-y|^p \, \gamma(dx, dy) \right)^{\frac{1}{p}},$$

where $|\cdot|$ is the Euclidean norm in $\mathbf{R}^d$ and $\Gamma(\mu, \nu)$ denotes the set of all couplings of μ and ν, namely,

$$\gamma(dx, \mathbf{R}^d) = \mu(dx) \quad \text{and} \quad \gamma(\mathbf{R}^d, dy) = \nu(dy)$$

In general, for two empirical measures of same size $\mu_N := \frac{1}{N}\sum_{i=1}^N \delta_{U_i}$ and $\nu_N := \frac{1}{N}\sum_{i=1}^N \delta_{V_i}$, it is well-known that

$$\mathcal{W}_p(\mu_N, \nu_N) := \left(\inf_{\sigma \in \mathcal{S}([N])} \frac{1}{N} \sum_{i=1}^N \left| U_i - V_{\sigma(i)} \right|^p \right)^{\frac{1}{p}},$$

where σ denotes a permutation in $\mathcal{S}([N])$, which is the collection of all the permutations on the set $[N] := \{1, 2, \ldots, N\}$. The proof will be provided later. (In fact, it is immediate by considering the original Kantorovich's formulation of Optimal Transport. see, e.g. [AG13].)

Now, we consider the Wasserstein distance between two empirical measures with different size, say, $\mu_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{U_i}$ and $\nu_M = \frac{1}{M}\sum_{i=1}^{M}\delta_{V_i}$. By considering the following alternative forms

$$\mu_N = \frac{1}{MN}\sum_{i=1}^{MN}\delta_{U_{\lceil i/M\rceil}} \quad \text{and} \quad \nu_M = \frac{1}{MN}\sum_{i=1}^{MN}\delta_{V_{\lceil i/N\rceil}},$$

we deduce that

$$\mathcal{W}_p(\mu_N,\nu_M) := \left(\inf_{\sigma\in\mathcal{S}([MN])}\frac{1}{MN}\sum_{i=1}^{MN}\left|U_{\lceil i/M\rceil} - V_{\lceil\sigma(i)/N\rceil}\right|^p\right)^{\frac{1}{p}}.$$

Interestingly, when $M = 1$, we have

$$\begin{aligned}\mathcal{W}_p(\mu_N,\nu_1) &:= \left(\inf_{\sigma\in\mathcal{S}([N])}\frac{1}{N}\sum_{i=1}^{N}\left|U_{\lceil i\rceil} - V_{\lceil\sigma(i)/N\rceil)}\right|^p\right)^{\frac{1}{p}}\\ &= \left(\frac{1}{N}\sum_{i=1}^{N}\left|U_{\lceil i\rceil} - V_{\lceil i/N\rceil)}\right|^p\right)^{\frac{1}{p}}\\ &= \left(\frac{1}{N}\sum_{i=1}^{N}|U_i - V_1|^p\right)^{\frac{1}{p}}.\end{aligned}$$

In particular, we have

$$\mathcal{W}_2\left(\delta_{Y_i},\frac{1}{N}\sum_{j=1}^{N}\delta_{Y_j}\right)^2 = \frac{1}{N}\sum_{j=1}^{N}(Y_i - Y_j)^2, \tag{1}$$

whence

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathcal{W}_2\left(\delta_{Y_i},\frac{1}{N}\sum_{j=1}^{N}\delta_{Y_j}\right)^2 &= \frac{1}{N^2}\sum_{j=1}^{N}\sum_{j=1}^{N}(Y_i - Y_j)^2\\ &= \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y}_N)^2.\end{aligned} \tag{2}$$

4 Splitting criteria

In this section, we take a closer look at the design of the splitting criteria.

4.1 Breiman's splitting criteria

Thanks to (2), the Breiman's splitting criteria w.r.t. A_L and A_R can be reformulated as the following form

$$\begin{aligned}&L_{reg}(A_L,A_R)\\ :=&\frac{1}{N_A}\sum_{i=1}^{N_A}(Y_i - \bar{Y}_A)^2 - \frac{1}{N_A}\sum_{X_i\in A_L}(Y_i - \bar{Y}_{A_L})^2 - \frac{1}{N_A}\sum_{X_i\in A_R}(Y_i - \bar{Y}_{A_R})^2.\\ =&\frac{1}{N_A}\sum_{i=1}^{N_A}(Y_i - \bar{Y}_A)^2 - \frac{N_L}{N_A}\left(\frac{1}{N_L}\sum_{X_i\in A_L}(Y_i - \bar{Y}_{A_L})^2\right) - \frac{N_R}{N_A}\left(\frac{1}{N_R}\sum_{X_i\in A_R}(Y_i - \bar{Y}_{A_R})^2\right)\\ =&\frac{1}{N_A}\sum_{i=1}^{N_A}\mathcal{W}_2\left(\delta_{Y_i},\pi_A\right)^2 - \frac{N_L}{N_A}\left(\frac{1}{N_L}\sum_{X_i\in A_L}\mathcal{W}_2\left(\delta_{Y_i},\pi_L\right)^2\right) - \frac{N_R}{N_A}\left(\frac{1}{N_R}\sum_{X_i\in A_R}\mathcal{W}_2\left(\delta_{Y_i},\pi_R\right)^2\right).\end{aligned} \tag{3}$$

To give some intuition, imagine that the **nodesize** is set to be 1, namely, for each decision tree, we only use one point to provide the prediction. In such a scenario, what we should minimize is the Wassertein distance between a singular empirical measure δ_{Y_i} with the conditional distribution. Instead of minimizing the $\mathcal{W}_2$-distance, we minimize the *square* of the $\mathcal{W}_2$-distance. The Breiman's criterion can thus be interpreted by maximizing the gain of average of the objective, i.e., the *square* of $\mathcal{W}_2$-distance induced by the current split. In fact, without going further, we have already an idea of how to split in the multivariate case: we use the interpretation of $\mathcal{W}_2$-distance, which does not depend on the underlying dimension of Y. The complexity is linear w.r.t. the dimension of Y.

Another interesting remark is that

$$\mathcal{W}_2\left(\frac{\delta_{Y_1}+\delta_{Y_2}}{2},\pi_A\right)^2 \leq \frac{1}{2}\left(\mathcal{W}_2(\delta_{Y_1},\pi_A)^2+\mathcal{W}_2(\delta_{Y_2},\pi_A)^2\right).$$

In fact, by definition

$$\mathcal{W}_2\left(\frac{\delta_{Y_1}+\delta_{Y_2}}{2},\pi_A\right)^2 = \inf_{\sigma\in\mathcal{S}([2N])}\frac{1}{2N}\sum_{i=1}^{N}\left|Y_{\lceil i/N\rceil}-Y_{\lceil\sigma(i)/2\rceil)}\right|^2.$$

Fixing $\sigma\in\mathcal{S}([2N])$, we consider

$$\mathcal{S}([N])\ni\sigma_1: i\mapsto\lceil\sigma(i)/2\rceil) \quad \text{and} \quad \mathcal{S}([N])\ni\sigma_2: i\mapsto\lceil\sigma(i+N)/2\rceil).$$

Then, it is easy to check that

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^{2N}\left|Y_{\lceil i/N\rceil}-Y_{\lceil\sigma(i)/2\rceil)}\right|^2\\
\leq&\frac{1}{2N}\sum_{i=1}^{N}\left|Y_1-Y_{\lceil\sigma_1(i)\rceil)}\right|^2+\frac{1}{2N}\sum_{i=1}^{N}\left|Y_2-Y_{\lceil\sigma_2(i)\rceil)}\right|^2\\
=&\frac{1}{2}\left(\mathcal{W}_2(\delta_{Y_1},\pi_A)^2+\mathcal{W}_2(\delta_{Y_2},\pi_A)^2\right).
\end{aligned}$$

Multatis mutantis, it is clear that

$$\mathcal{W}_p\left(\frac{1}{\#(I)}\sum_{i\in I}\delta_{Y_i},\pi_A\right)^p \leq \frac{1}{\#(I)}\sum_{i\in I}\mathcal{W}_p(\delta_{Y_i},\pi_A)^p,$$

for any index set $I\subset[N]$. From this point of view, minimizing the average of $\mathcal{W}_p^p$ between the singular measures and a reference empirical measure can be regarded as minimizing an upper bound of $\mathcal{W}_p^p$ between the associated empirical measure and the reference empirical measure.

4.2 Wasserstein splitting criteria

The original idea of introducing the splitting that is compatible with Wasserstein distance is to maximize $\mathcal{W}_p(\pi_L,\pi_R)$. However, as is mentioned in the last document, this strategy failed to provide reliable estimation in high dimensional setting. For example, it does not work in the case where X contains a lot of irrelevant features. This is partially due to the fact that this splitting criterion does not take into account the number of data points: the splits turn out to be conducted much more frequently in the region close to the boundary of the current node, since the Wasserstein distance tends to be large when the sizes of empirical measures differ greatly. Hence, a natural way to design a splitting criterion that

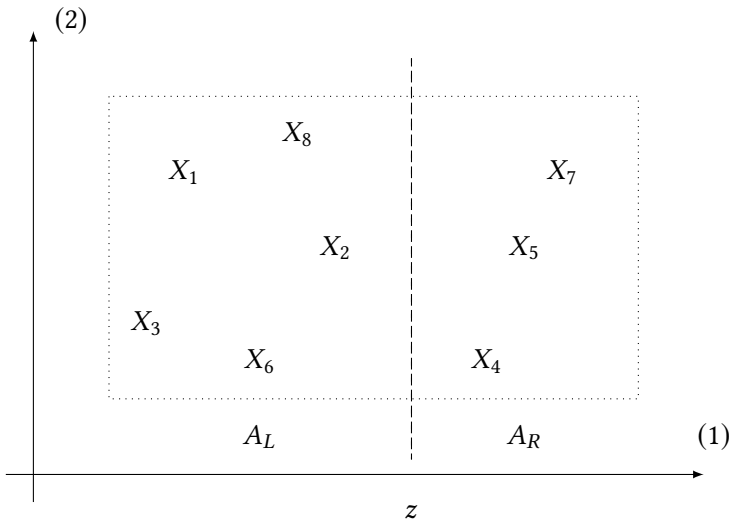

Figure 1: An illustration of splitting.

depends on the size of empirical measures is to weight them w.r.t. the number of data points therein. We consider the following splitting criteria

$$L_{\mathcal{W}_p}(A_L, A_R) := \left\{ \frac{N_L}{N_A} \mathcal{W}_p(\pi_L, \pi_A) + \frac{N_R}{N_A} \mathcal{W}_p(\pi_R, \pi_A) \right\}.$$

Note that the splitting z is always performed in the middle of two consecutive data points along the direction (1) in order to remove the possible ties in the arg max.
More concretely, in Figure 1 the Wasserstein distance to be evaluated is

$$\frac{5}{8}\mathcal{W}_p\left(\frac{1}{5}(\delta_{Y_1} + \delta_{Y_2} + \delta_{Y_3} + \delta_{Y_6} + \delta_{Y_8}), \frac{1}{8}\sum_{i=1}^{8}\delta_{Y_i}\right) + \frac{3}{8}\mathcal{W}_p\left(\frac{1}{3}(\delta_{Y_4} + \delta_{Y_5} + \delta_{Y_7}), \frac{1}{8}\sum_{i=1}^{8}\delta_{Y_i}\right)$$

The idea behind remains the same: the split is done in order to maximize the homogeneity between the nodes.

4.3 Connection between these criteria

There is an interesting connection between the Breiman's criterion and Wasserstein criterion. Let us consider the generalized Breiman-Wasserstein criterion defined as follows:

$$\begin{aligned} &L_{reg}^{\mathcal{W}_p}(A_L, A_R) \\ =&\frac{1}{N_A}\sum_{i=1}^{N_A}\mathcal{W}_p\left(\delta_{Y_i}, \pi_A\right)^p - \frac{N_L}{N_A}\left(\frac{1}{N_L}\sum_{X_i \in A_L}\mathcal{W}_p\left(\delta_{Y_i}, \pi_L\right)^p\right) - \frac{N_R}{N_A}\left(\frac{1}{N_R}\sum_{X_i \in A_R}\mathcal{W}_p\left(\delta_{Y_i}, \pi_R\right)^p\right), \end{aligned} \tag{4}$$

which in particular coincides with Breiman's criteria when $p = 2$ and the underlying dimension of Y is 1. Let us see the case where $p = 1$.

$$\begin{aligned} &L_{reg}^{\mathcal{W}_1}(A_L, A_R) \\ =&\frac{1}{N_A}\sum_{i=1}^{N_A}\mathcal{W}_1\left(\delta_{Y_i}, \pi_A\right) - \frac{N_L}{N_A}\left(\frac{1}{N_L}\sum_{X_i \in A_L}\mathcal{W}_1\left(\delta_{Y_i}, \pi_L\right)\right) - \frac{N_R}{N_A}\left(\frac{1}{N_R}\sum_{X_i \in A_R}\mathcal{W}_1\left(\delta_{Y_i}, \pi_R\right)\right) \\ =&\frac{1}{N_A}\sum_{X_i \in A_L}\left(\mathcal{W}_1(\delta_{Y_i}, \pi_A) - \mathcal{W}_1(\delta_{Y_i}, \pi_L)\right) + \frac{1}{N_A}\sum_{X_i \in A_R}\left(\mathcal{W}_1(\delta_{Y_i}, \pi_A) - \mathcal{W}_1(\delta_{Y_i}, \pi_R)\right). \end{aligned} \tag{5}$$

Thanks to triangle inequality, we have

$$\mathcal{W}_1(\delta_{Y_i}, \pi_A) - \mathcal{W}_1(\delta_{Y_i}, \pi_L) \leq \mathcal{W}_1(\pi_A, \pi_L),$$

as well as

$$\mathcal{W}_1(\delta_{Y_i}, \pi_A) - \mathcal{W}_1(\delta_{Y_i}, \pi_R) \leq \mathcal{W}_1(\pi_A, \pi_R),$$

which yields

$$L_{\mathcal{W}_1}(A_L, A_R) \leq L_{reg}^{\mathcal{W}_1}(A_L, A_R).$$

Remark 4.1. I am still struggling to understand all these connections, which seems to be extremely interesting. For the numerical test on $\mathcal{W}_2$-distance, the best player is currently $L_{\mathcal{W}_2}$, which is just slightly better than $L_{reg}^{\mathcal{W}_2}$. However, the complexity of splitting is $\mathcal{O}(N \log N)$. In addition, the generalization to the multivariate case of $L_{reg}^{\mathcal{W}_2}$ is almost free. It is also important to remark that $p = 2$ seems to be the single setting such that a $\mathcal{O}(N)$ splitting is available, even in multivariate case. I will look into this later and a detailed numerical report should be available hopefully next week.

References

[AG13] Luigi Ambrosio and Nicola Gigli. *A User's Guide to Optimal Transport*, pages 1–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.