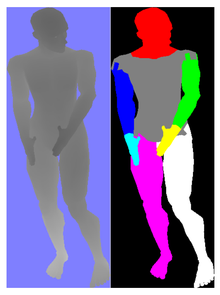
Pruning randomized trees with *L1*-norm regularization

Arnaud Joly, François Schnitzler, Pierre Geurts, Louis Wehenkel
Systems and modeling, Department of EE and CS, University of Liège

Motivation

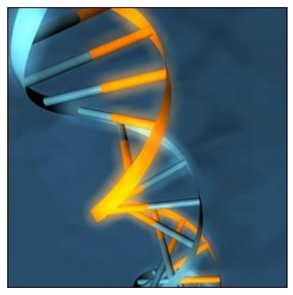
Goal of Supervised learning: from a dataset of input-output pairs $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$ to learn a function $f : \mathcal{X} \to \mathcal{Y}$ to predict the output y for any (new) input x.

Supervised learning is often applied on high-dimensional data:



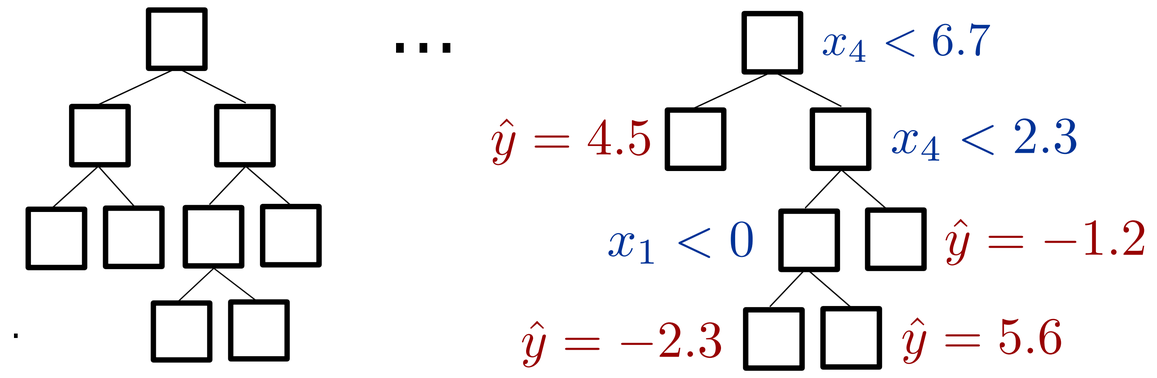
3D Image segmentation (Kinect)

Stability prediction in electrical network



Genome studies

A robust supervised learning method: ensemble of randomized decision trees (eg., extremely randomized trees , Geurts et al., 2006). Each decision tree is a hierarchical set of questions which leads to a prediction.



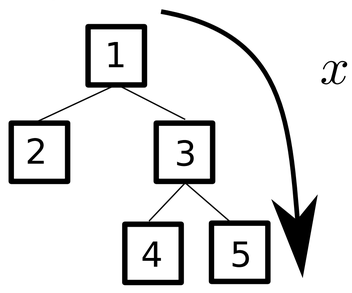
High dimensional data often requires huge ensembles of randomized trees to achieve good performance, i.e.
- a high number of trees,
- each tree is fully developed (no pruning)

→ **Huge amount of storage**

Method: pruning with L1-regularization

Each tree node N is associated with a prediction weight w and a node characteristic function $I(x \in N)$ equals to 1 if the sample (x, y) reaches the node, 0 otherwise.

Example for one tree and one sample:



$$W = [0, w_2, 0, w_4, w_5]^T$$
$$A(x) = [I(x \in N_1), \ldots, I(x \in N_5)]^T$$
$$\hat{f}(x) = W^T A(x) = w_5$$

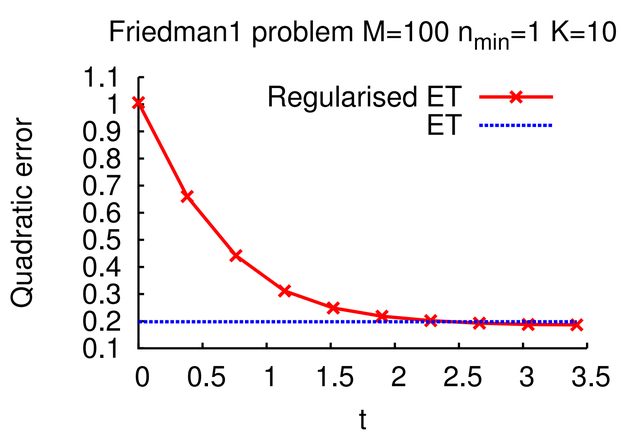
The function $A(x) = [I(x \in N_1), \ldots, I(x \in N_q)]^T$ thus embeds the original input space $\mathcal{X}$ into a new space $\mathcal{Z} \subseteq \{0, 1\}^q$ with q the total number of tree nodes.

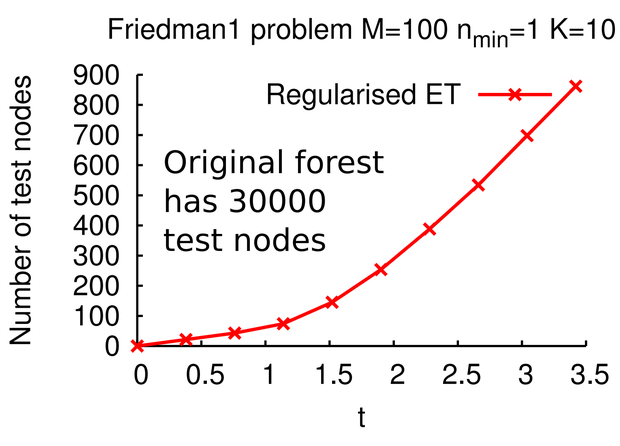
Variable selection is performed by applying L1-norm regularization (LASSO) on the new training samples $\{(z_i = (z_{i,1}, \ldots, z_{i,q}), y_i)\}_{i=1}^n \subseteq \mathcal{Z} \times \mathcal{Y}$:

$$\min_{\{\beta_0, \ldots, \beta_q\}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{q} \beta_j \, z_{ij} \right)^2$$
$$s.t. \qquad \sum_{j=1}^{q} |\beta_j| \leq t.$$

The solution $\{\beta_j\}_{j=1}^q$ leads to prune and to reweight the randomized tree ensemble. A test node can be deleted if the weights β_j of all its descendants are equal to zero.

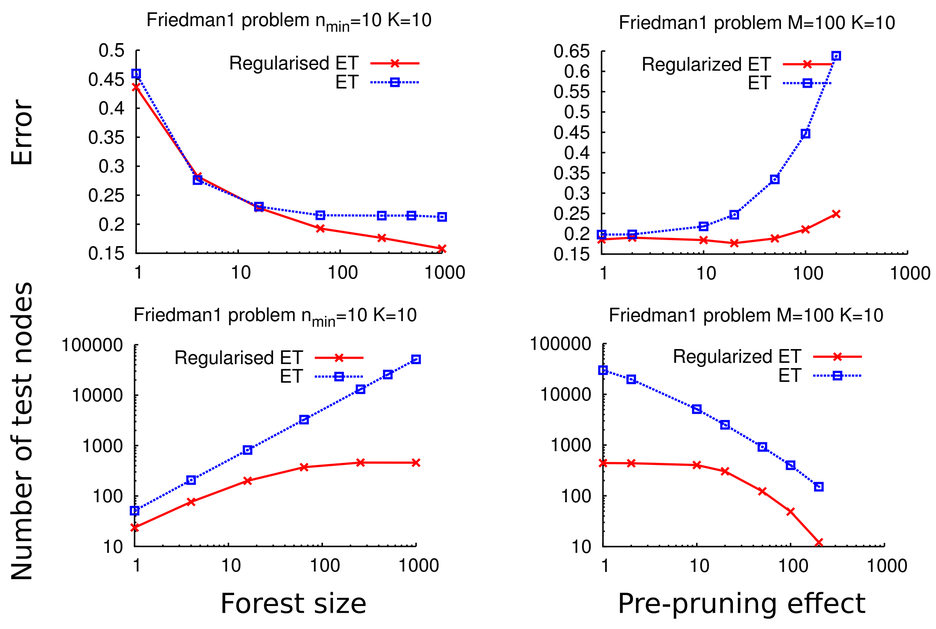
Results





Regularization preserves accuracy or even can improve it.

Drastic pruning.



Robust with respect to the number of trees and the pre-pruning parameter

Conclusion and perspectives

Our experiments show that it is possible to drastically prune ensembles of randomized trees while preserving or even improving their accuracy.

Future research:
- how to avoid the generation of a huge randomized tree ensemble prior to regularization?
- Study the links with compressed sensing

Interested? Email me at a.joly@ulg.ac.be or go on www.ajoly.org