General response. First of all, we thank the reviewers for their helpful comments and remarks. We provide first general comments prior to address additional points raised by the reviewers.

We want to stress that we have proposed a generalization of regression trees that (1) adapt to the smoothness of the prediction function relating input and output variables while (2) preserving the interpretability of the prediction and (3) being robust to noise. The three points, smoothness, interpretability and robustness to noise, are all important and have been illustrated empirically. There is however *no free lunch*, and these additional properties come with a computational cost, as described for training in Appendix A.2 (note that, as mentioned in the main paper line 193, we make use of the Moore-Penrose pseudo-inverse which explains why the complexity is only quadratic in K). Applying PR trees is also more costly than applying standard decision trees as the function Ψ (Eq. 2) needs to be evaluated on all regions. We provide below the prediction time, in seconds, on some datasets (we'll include these results in Appendix A.2).

Dataset	PR Tree	Std Tree	# Observations
BD	0.3	1.00E-04	146
BO	0.24	1.00E-04	101
DI	0.2	9.00E-05	88
RI	0.04	2.00E-04	14

We also want to emphasize that the theoretical framework we propose does not assume that the $x_j, 1 \leq j \leq p$, are independent. The notation $\phi\left(()_{1\leq j\leq p}\right)$ in Eq. 3 means that ϕ is a multivariate function of the p variables $\frac{u_j - x_j}{\sigma_j}$. For convenience, we have used functions ϕ that lead to standard cdfs for Ψ in our experiments, dropping the dependencies between x_j. Other choices could be made, in particular when dependencies between x_j are known. In any case, u_j cannot be interpreted as a location parameter or as a center of a region as it is the variable that is integrated out.

Reviewer 1. Ooops, you are right: The expectation in the expression of a_n in Proposition 1 should be removed (this proposition directly derives from Proposition C2 in Appendix C2.2, with no expectation; the expectation should also be removed from Proposition C3 in Appendix 2.4). The regions $\mathcal{R}_k^{(n)}$ are fixed for a given n.

Reviewer 2. An important difference wrt to the work by Gérard Biau, Luc Devroye and Gabor Lugosi ([1]) is that we are not averaging over independent classifiers as regions are dependent on each other. Our consistency proof radically differs from theirs because of this difference.

Adaptative Neural Trees ([2]) and Deep Neural Decision Forests ([3]) are both built from decision trees. These models are very close to soft trees, to which we compare ourselves. In each case however, the models are enhanced with a neural network representation and suffer from a lack of interpretability (one can even argue that these models are not tree models *per se*). The paper of Forsst & Hinton ([4]) considers a specific variant of the soft tree model, with knowledge distillation. Distilling knowledge into our trees is clearly an interesting research direction that we plan to investigate.

Because of their interpretability, decision trees seem to be still heavily used in the industry, as mentioned in the 2019 Kaggle survey (https://www.kaggle.com/kaggle-survey-2019). This said, Random Forests aim at reducing the variance (and this comes at the expense of a small increase in the bias) whereas our adaptation to smoothness aims at reducing the bias. Combining both, as in PR-RF, reduces both bias and variance and leads to a method which significantly outperforms RF (Table 5, Appendix A.4).

Reviewer 3. It is true that a standard regression tree with enough leaves can also approximate a smooth link function. However, to obtain such a tree, one needs large samples, which are unfortunately not available in practice (as examplified, *e.g.*, by the difference between standard and PR trees in our experiments).

Uncertain decision trees were designed to deal with uncertainty in the input variables and rely on a set of given pdfs modeling the uncertainty on each attribute value for this particular example. This contrats with our approach that aims at adapting to the smoothness of the prediction function. In particular, the intervals $[a_{i,j}, b_{i,j}]$ (reference [24] of our paper) defining the support of the pdfs are given in uncertain decision trees whereas they are learned in our case.

Our discussion on overfitting simply amounts to saying that the more complex a model is, the more likely it is to overfit (in practice, the amount of samples available is usually not large enough to avoid that). We'll modify lines 228-230 as we agree that they may be confusing. The additional complexity of PR trees compared to standard trees is not important and has not led to overfitting in our experiments.

There is a typo in line 194 as it is Ψ (and not ϕ) that corresponds to the cdf of a normal distribution (multivariate normal distribution with diagonal covariance matrix equal to $\boldsymbol{\sigma}$).

Reviewer 4. One can obtain standard regression trees from Eqs 2 and 3 by setting ϕ to $(2\pi)^{-\frac{p}{2}} \prod_{j=1}^{p} \exp(-\frac{(u_j - x_j)^2}{2\sigma_j^2})$, with $\sigma_j \to 0$ for all j. In that case, the distribution of $\mathbf{x}$ over regions is concentrated on one region.