computation

Article

Extreme Multiclass Classification Criteria

Anna Choromanska * and Ish Kumar Jain

NYU Tandon School of Engineering, Department of Electrical and Computer Engineering 5 MetroTech Center, Brooklyn, NY 11201, USA; ishjain@nyu.edu

* Correspondence: ac5455@nyu.edu

Received: 2 February 2019; Accepted: 8 March 2019; Published: 12 March 2019

Abstract: We analyze the theoretical properties of the recently proposed objective function for efficient online construction and training of multiclass classification trees in the settings where the label space is very large. We show the important properties of this objective and provide a complete proof that maximizing it simultaneously encourages balanced trees and improves the purity of the class distributions at subsequent levels in the tree. We further explore its connection to the three well-known entropy-based decision tree criteria, i.e., Shannon entropy, Gini-entropy and its modified variant, for which efficient optimization strategies are largely unknown in the extreme multiclass setting. We show theoretically that this objective can be viewed as a surrogate function for all of these entropy criteria and that maximizing it indirectly optimizes them as well. We derive boosting guarantees and obtain a closed-form expression for the number of iterations needed to reduce the considered entropy criteria below an arbitrary threshold. The obtained theorem relies on a weak hypothesis assumption that directly depends on the considered objective function. Finally, we prove that optimizing the objective directly reduces the multi-class classification error of the decision tree.

Keywords: multiclass classification; decision trees; boosting

1. Introduction

This paper focuses on the multiclass classification setting, where the number of classes is very large. The recent widespread development of data-acquisition web services and devices has helped make large data sets, such as multiclass data sets, commonplace. Straightforward extensions of the binary approaches to the multiclass setting, such as the one-against-all approach [1], which for each data point computes a score for each class and returns the class with the maximum score, do not often work in the presence of strict computational constraints as their running time often scales linearly with the number of labels k. On the other hand, the most computationally efficient approaches for multiclass classification are given by $\mathcal{O}(\log k)$ train/test running time [2]. This running time can naturally be achieved by hierarchical classifiers that build the hierarchy over the labels.

This paper considers a hierarchical multiclass decision tree structure, where each node of the tree contains a binary classifier h from some hypothesis class $\mathcal{H}$ that sends an example reaching that node to either left ($h(x) \leq 0$) or right ($h(x) > 0$) child node depending on the sign of $h(x)$ (each node has its own splitting hypothesis). The test example descends from the root to the leaf of such tree guided by the classifiers lying on its path, and is labeled according to the label with the highest frequency amongst the training examples that were reaching the leaf that it descended to. The tree is constructed and trained in a top-down fashion, where splitting the data in every node of the tree is done by maximizing the following objective function recently introduced in the literature [3] (along with the algorithm (we refer the reader to the referenced paper for the algorithm's details), called LOMtree, optimizing it in an online fashion):

$$J(h) := 2\sum_{i=1}^{k} |\pi_i P(h(x) > 0) - \underbrace{P(h(x) > 0, i)}_{P(h(x)>0|i)\pi_i}|, \tag{1}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}^d$ are the data points (each with a label from the set $\{1, 2, \ldots, k\}$), π_i denotes the proportion of label i amongst the examples reaching a node, and probabilities $P(h(x) > 0)$ and $P(h(x) > 0|i)$ denote the fraction of examples reaching a node for which $h(x) > 0$, marginally and conditional on class i respectively. The objective measures the dependence between the split and the class distribution. Note that it satisfies $J(h) \in [0, 1]$ and, as implied by its form, maximizing it encourages the fraction of examples going to the right from class i to be substantially different from the background fraction for each class i. Thus for a balanced split (i.e., $P(h(x) > 0) = 0.5$), the examples of class i are encouraged to be sent exclusively to the left ($P(h(x) > 0|i) = 0$) or right ($P(h(x) > 0|i) = 1$) refining the purity of the class distributions at subsequent levels in the tree. The LOMtree algorithm effectively maximizes this objective over hypotheses $h \in \mathcal{H}$ in an online fashion with stochastic gradient descent (SGD) and obtains good-quality multiclass tree predictors with logarithmic train and test running times. Despite that, this objective and its properties (including the relation to the more standard entropy criteria) remain largely ununderstood. Its exhaustive analysis is instead provided in this paper.

Our contributions are the following:

- We provide an extensive theoretical analysis of the properties of the considered objective and prove that maximizing this objective in any tree node simultaneously encourages balanced partition of the data in that node and improves the purity of the class distributions at its children nodes.
- We show a formal relation of this objective to some more standard entropy-based objectives, i.e., Shannon entropy, Gini-entropy and its modified variant, for which online optimization schemes in the context of multiclass classification are largely unknown. In particular we show that i) the improvement in the value of entropy resulting from performing the node split is lower-bounded by an expression that increases with the value of the objective and thus ii) the considered objective can be used as a surrogate function for indirectly optimizing any of the three considered entropy-based criteria.
- We present three boosting theorems for each of the three entropy criteria, which provide the number of iterations needed to reduce each of them below an arbitrary threshold. Their weak hypothesis assumptions rely on the considered objective function.
- We establish the error bound that relates maximizing the objective function with reducing the multi-class classification error.
- Finally, in the Appendix A we establish an empirical connection between the multiclass classification error and the entropy criteria and show that Gini-entropy most closely resembles the behavior of the test error in practice.

The main theoretical analysis of this paper is kept in the boosting framework [4] and relies on the assumption that the objective function can be weakly optimized in the internal nodes of the tree. This weak advantage is amplified in the tree leading to hierarchies achieving any desired level of entropy (either Shannon entropy, Gini-entropy or its modified variant). Our work adds new theoretical results to the theory of multiclass boosting. Note that the multiclass boosting is largely ununderstood from the theoretical perspective [5] (we refer the reader to [5] for comprehensive review of the theory of muticlass boosting).

The paper is organized as follows: related literature is discussed in Section 2, the theoretical properties of the objective $J(h)$ are shown in Section 3, the main theoretical results are presented in Section 4, and finally the mathematical properties of the entropy criteria and the proofs of the main theoretical results are provided in Section 5. Conclusions (Section 6) end the paper. Appendix A contains basic numerical experiments (Appendix A.1) and additional proofs (Appendix A.2).

2. Related Work

The extreme multiclass classification problem has been addressed in the literature in different ways. We discuss them here, putting emphasis on the ones that build hierarchical predictors as these techniques are the most relevant to this paper. Only a few authors [2,3,6–8] simultaneously address logarithmic time training and testing. The methods they propose are either hard to apply in practical problems [7] or use fixed tree structures [6,8]. Furthermore, an alternative approach based on using a random tree structure was shown to potentially lead to considerable underperformance [3,9]. At the same time, for massive datasets making multiple passes through the data is computationally costly, which justifies the need for developing online approaches, where the algorithm streams over a potentially infinitely large data set (online approaches are also plausible for non-stationary problems). It is unclear how to optimize standard decision tree objectives, such as Shannon or Gini-entropy, in this setting (early attempt was recently proposed [2] for Shannon entropy). One of the prior works to this paper [3] introduces an objective function which enjoys certain advantages over entropy criteria. In particular, it can be easily and efficiently optimized online. The authors however present an incomplete theoretical analysis and leave a number of open questions, which this paper instead aims at addressing. The algorithms for incremental learning of classification with decision trees also include some older works [10–12], which split any node according to the outcome of the node split-test based on the values of selected attributes of the data examples reaching that node. These approaches are different from the one in this paper, where the node split is performed according to the value of the learned (e.g., with SGD) hypothesis computed for the entire vector of attributes of the data examples reaching that node.

Other tree-based approaches include conditional probability trees [13] and clustering methods [9,14,15] ([9] was later improved in [16]), but they allow training time to be linear in the label complexity. The remaining techniques for multiclass classification include sparse output coding [17], variants of error correcting output codes [18], variants of iterative least-squares [19], and a method based on guess-averse loss functions [20].

Finally note that the conditional density estimation problem is also challenging in the large-class settings and in this respect remains parallel to the extreme multiclass classification problem [21]. In the context of conditional density estimation problem, there have also been some works that use tree structured models to accelerate computation of the likelihood and gradients [8,22–24]. They typically use heuristics based on using ontologies [8], Huffman coding [24], and various other mechanisms.

3. Theoretical Properties of the Objective Function

In this section we describe the objective function introduced in Equation (1) and provide its theoretical properties. The proofs are deferred to the Appendix. We first introduce the definitions of the concept of *balancedness* and *purity* of the node split.

Definition 1 (Purity and balancedness)**.** *The hypothesis $h \in \mathcal{H}$ induces a pure split if $\alpha := \sum_{i=1}^{k} \pi_i \min(P(h(x) > 0|i), P(h(x) < 0|i)) \leq \delta$, where $\delta \in [0, 0.5)$, and α is called the purity factor.*

The hypothesis $h \in \mathcal{H}$ induces a balanced split if $\beta := P(h(x) > 0) \in [c, 1-c]$, where $c \in (0, 0.5]$, and β is called the balancing factor.

A partition is *perfectly pure* if $\alpha = 0$ (examples of the same class are sent exclusively to the left or to the right). A partition is called *perfectly balanced* if $\beta = 0.5$ (equal number of examples are sent to the left and to the right). The notions of balancedness and purity are conveniently illustrated in Figure 1, where it is shown that the purity criterion helps to refine the choice of the splitting hypothesis from among well-balanced candidates.

Next, we show the first theoretical property of the objective function $J(h)$ that characterizes its behavior at the optimum ($J(h) = 1$).

Lemma 1. *The hypothesis $h \in \mathcal{H}$ induces a perfectly pure and balanced partition if and only if $J(h) = 1$.*

For some data sets however there exist no hypotheses producing perfectly pure and balanced splits. We next show that increasing the value of the objective leads to more balanced splits.

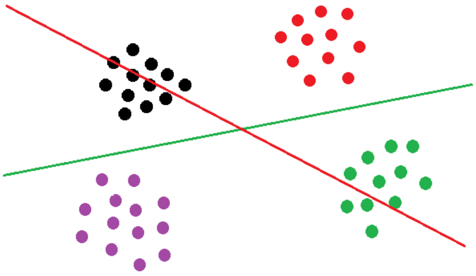

Figure 1. **Red partition**: highly balanced split but impure (the partition cuts through the black and green classes). **Green partition**: highly balanced and highly pure split. Figure should be read in color.

Lemma 2. *For any hypothesis h and any distribution over data examples the balancing factor β satisfies $\beta \in \left[0.5(1 - \sqrt{1 - J(h)}), 0.5(1 + \sqrt{1 - J(h)})\right]$.*

We refer to the interval to which β belongs to as β-interval. Thus the larger (closer to 1) the value of $J(h)$ is, the narrower the β-interval is, leading to more balanced splits at the extremes of this interval (β closer to 0.5).

This result combined with the next lemma implies that, at the extremes of the β interval, the value of the upper-bound on the purity factor decreases as the value of $J(h)$ increases (since $J(h)$ gets closer to 1 and the balancing factor β gets closer to 0.5 at the extremes of the β interval). The recovered splits therefore have better purity (α closer to 0).

Lemma 3 (Lemma 1 in [3])**.** *For any hypothesis h and any distribution over data examples the purity factor α and the balancing factor β satisfy $\alpha \leq \min\{(2 - J(h))/4\beta - \beta, 0.5\}$.*

Note that the equality condition in Lemma 3 is achieved when $P(h(x) > 0|i) = P(h(x) < 0|i) = 0.5$ (and thus, $\alpha = 0$, $\beta = 0.5$, and $J(h) = 0$).

We thus showed that maximizing the objective in Equation (1) in each tree node simultaneously encourages trees that are balanced and whose purity of the class distributions is gradually improving when moving from the root to a subsequent tree levels. Lemmas 2 and 3 are illustrated in Figure 2.

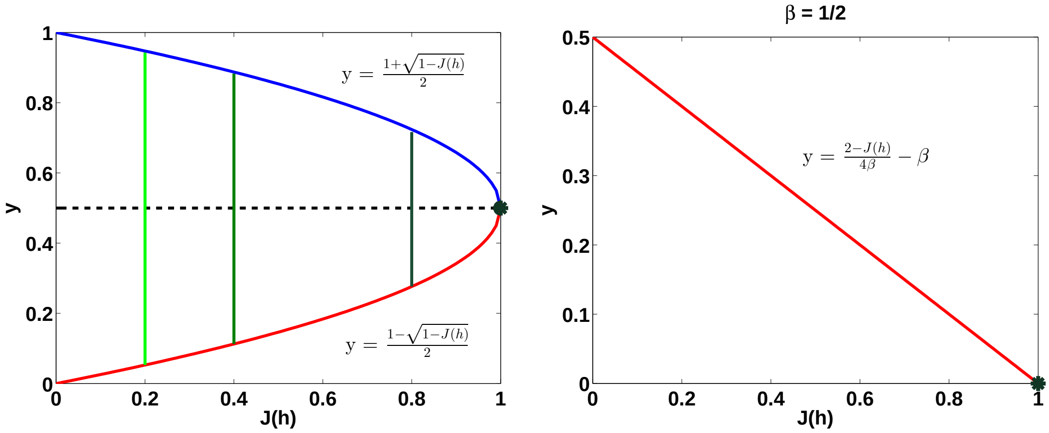

Figure 2. **Left**: Blue curve captures the behavior of the upper-bound on the balancing factor as a function of $J(h)$, red curve captures the behavior of the lower-bound on the balancing factor as a function of $J(h)$, green intervals correspond to the intervals where the balancing factor lies for different values of $J(h)$. **Right**: Red line captures the behavior of the upper-bound on the purity factor as a function of $J(h)$ when the balancing factor is fixed to $\frac{1}{2}$. Figure should be read in color.

In the next section we show that the objective $J(h)$ is related to the more standard decision tree entropy-based objectives and that maximizing it leads to the reduction of these criteria. We consider three different entropy criteria in this paper. The theoretical analysis relies on the boosting framework and depends on the weak learning assumption. Three different entropy-based criteria lead to three different theoretical statements, where we bound the number of splits required to reduce the value of the criterion below given level. The bounds we obtain, and their dependences on the number of classes (k), critically depend on the strong concativity properties of the considered entropy-based objectives.

4. Main Theoretical Results

4.1. Notation

We first introduce notation. Let $\mathcal{T}$ denote the tree under consideration. $\pi_{l,i}$'s denote the probabilities that a randomly chosen data point x drawn from $\mathcal{P}$, where $\mathcal{P}$ is a fixed target distribution over $\mathcal{X}$, has label i given that x reaches node l (note that $\sum_{i=1}^{k} \pi_{l,i} = 1$), t denotes the number of internal tree nodes, $\mathcal{L}_t$ denotes the set of all tree leaves at time t, and w_l is the weight of leaf l defined as the probability a randomly chosen x drawn from $\mathcal{P}$ reaches leaf l (note that $\sum_{l \in \mathcal{L}_t} w_l = 1$). We study a tree construction algorithm where we recursively find the leaf node with the highest weight, and choose to split it into two children. Consider the tree constructed over t steps where in each step we take one leaf node and split it (thus the number of splits is equal to the number of internal nodes of the tree) ($t = 1$ corresponds to splitting the root, thus the tree consists of one node (root) and its two children (leaves) in this step). We measure the quality of the tree at any given time t with three different entropy criteria:

- Shannon entropy G_t^e:

$$G_t^e = \sum_{l \in \mathcal{L}_t} w_l \sum_{i=1}^{k} \pi_{l,i} \ln\left(\frac{1}{\pi_{l,i}}\right)$$

- Gini-entropy G_t^g:

$$G_t^g = \sum_{l \in \mathcal{L}_t} w_l \sum_{i=1}^{k} \pi_{l,i}(1 - \pi_{l,i})$$

- Modified Gini-entropy G_t^m:

$$G_t^m = \sum_{l \in \mathcal{L}_t} w_l \sum_{i=1}^{k} \sqrt{\pi_{l,i}(\mathcal{C} - \pi_{l,i})},$$

where $\mathcal{C}$ is a constant such that $\mathcal{C} > 2$.

These criteria are the natural extensions of the criteria used in the context of binary classification [25] to the multiclass classification setting (note that there is more than one way of extending the entropy-based criteria from [25] to the multiclass classification setting, e.g., the modified Gini-entropy could as well be defined as $G_t^m = \sum_{l \in \mathcal{L}_t} w_l \sum_{i=1}^k \sqrt{\pi_{l,i}(\mathcal{C} - \pi_{l,i})}$, where $\mathcal{C} \in [1, 2]$. This and other extensions will be investigated in future works). We will next present the main results of this paper, which will be followed by their proofs. We begin with introducing the weak hypothesis assumption.

4.2. Theorems

Definition 2 (Weak Hypothesis Assumption)**.** *Let m denote any internal node of the tree $\mathcal{T}$, and let $\beta_m = P(h_m(x) > 0)$ and $P_{m,i} = P(h_m(x) > 0|i)$. Furthermore, let $\gamma \in \mathbb{R}^+$ be such that for all m, $\gamma \in (0, \min(\beta_m, 1 - \beta_m)]$. We say that the* weak hypothesis assumption *is satisfied when for any distribution $\mathcal{P}$ over $\mathcal{X}$ at each node m of the tree $\mathcal{T}$ there exists a hypothesis $h_m \in \mathcal{H}$ such that $J(h_m)/2 = \sum_{i=1}^k \pi_{m,i}|P_{m,i} - \beta_m| \geq \gamma$.*

The weak hypothesis assumption says that in every node of the tree we are able to recover a hypothesis from $\mathcal{H}$ which corresponds to the value of the objective that is above 0 (thus the corresponding split is "weakly" pure and "weakly" balanced).

Consider next any time t and let n be the heaviest leaf at time t that we split and its weight w_n be denoted by w for brevity. Similarly, let h denote the regressor at node n (shorthand for h_n). We denote the difference between the contribution of node n to the value of the entropy-based objectives in times t and $t+1$ as

$$\Delta_t^e := G_t^e - G_{t+1}^e; \qquad \Delta_t^g := G_t^g - G_{t+1}^g; \qquad \Delta_t^m := G_t^m - G_{t+1}^m.$$

Then the following lemma holds (the proof in provided in Section 5):

Lemma 4. *Under the Weak Hypothesis Assumption, the change in entropies occuring due to the node split can be bounded as*

$$\Delta_t^e \geq \frac{wJ(h)^2}{8(1-\gamma)^2}; \quad \Delta_t^g \geq \frac{wJ(h)^2}{4k(1-\gamma)^2}; \quad \Delta_t^m \geq \frac{(\mathcal{C}-2)^2}{\mathcal{C}^3} \cdot \frac{wJ(h)^2}{4k(1-\gamma)^2}.$$

Clearly, maximizing the objective $J(h)$ improves the entropy reduction. The considered objective can therefore be viewed as a surrogate function for indirectly optimizing any of the three considered entropy-based criteria, for which efficient online optimization strategies are largely unknown but highly desired in the multiclass classification setting. To be more specific, the standard packages for binary classification trees, such as CART [26] and C4.5 [27], require running a brute force search to find a partition at every node of the tree from a set of all possible partitions that leads to the biggest improvement of the entropy-based criterion of interest [25]. This is prohibitive in case of the multiclass problem. $J(h)$ however can be efficiently optimized with SGD instead.

We next state the three boosting theoretical results captured in Theorems 1–3. They guarantee that the top-down decision tree algorithm which optimizes $J(h)$ in each node will amplify the weak advantage, captured in the weak learning assumption, to build a tree achieving any desired level of entropy (either Shannon entropy, Gini-entropy or its modified variant).

Theorem 1. *Under the Weak Hypothesis Assumption, for any $\alpha \in [0, 2\ln k]$, to obtain $G_t^e \leq \alpha$ it suffices to make $t \geq \left(\frac{2\ln k}{\alpha}\right)^{\frac{4(1-\gamma)^2}{\gamma^2 \log_2 e} \ln k}$ splits.*

Theorem 2. *Under the Weak Hypothesis Assumption, for any $\alpha \in [0, 2\left(1-\frac{1}{k}\right)]$, to obtain $G_t^g \leq \alpha$ it suffices to make $t \geq \left(\frac{2\left(1-\frac{1}{k}\right)}{\alpha}\right)^{\frac{2(1-\gamma)^2}{\gamma^2 \log_2 e}(k-1)}$ splits.*

Theorem 3. *Under the Weak Hypothesis Assumption, for any $\alpha \in [\sqrt{\mathcal{C}-1}, 2\sqrt{k\mathcal{C}-1}]$, to obtain $G_t^m \leq \alpha$ it suffices to make $t \geq \left(\frac{2\sqrt{k\mathcal{C}-1}}{\alpha}\right)^{\frac{2(1-\gamma)^2\mathcal{C}^3}{\gamma^2(\mathcal{C}-2)^2 \log_2 e}k\sqrt{k\mathcal{C}-1}}$ splits.*

Finally, we provide the error guarantee in Theorem 4. Denote $y(x)$ to be a fixed target function with domain $\mathcal{X}$, which assigns the data point x to its label, and let $\mathcal{P}$ be a fixed target distribution over $\mathcal{X}$. Together y and $\mathcal{P}$ induce a distribution on labeled pairs $(x, y(x))$. Let $t(x)$ be the label assigned to data point x by the tree. We denote as $\epsilon(\mathcal{T})$ the error of tree $\mathcal{T}$, i.e., $\epsilon(\mathcal{T}) :=_{x\sim\mathcal{P}} \left[\sum_{i=1}[t(x)=i, y(x)\neq i]\right]$

Theorem 4. *Under the Weak Hypothesis Assumption, for any $\alpha \in [0,1]$, to obtain $\epsilon(\mathcal{T}) \leq \alpha$ it suffices to make $t \geq \left(\frac{2\ln k \log_2 e}{\alpha}\right)^{\frac{4(1-\gamma)^2}{\gamma^2 \log_2 e}\ln k}$ splits.*

Remark 1. *The main theorems show how fast the entropy criteria or the multi-class classification error drop as the tree grows and performs node splits. These statements therefore provide a platform for comparing different entropy criteria and answer two questions: 1) for a fixed $\alpha, \gamma, \mathcal{C}$, and k, which criterion is reduced the most with each split? and 2) can the multi-class error match the convergence speed of the best entropic criterion? Hence, it can be noted that the Shannon entropy has the most advantageous dependence on the label complexity, since the bound scales only logarithmically with k, and thus achieves the fastest convergence. Simultaneously, the multi-class classification rate matches this advantageous convergence rate and also scales favorably (logarithmically) with k. Finally, even though the weak hypothesis requires only slightly favorable γ, i.e., $\gamma > 0$, in practice when constructing the tree one can optimize J in every node of the tree, which effectively pushes γ to be as high as possible. In that case γ becomes a well-behaving constant in the above theorems, ideally equal to $1/2$, and does not negatively affect the split count.*

We next discuss in details the mathematical properties of the entropy-based criteria, which are important to prove the above theorems.

5. Proofs

5.1. Properties of the Entropy-Based Criteria

Each of the presented entropy-based criteria has a number of useful properties that we give next, along with their proofs. We first give bounds on the values of the entropy-based functions. As before, let w be the weight of the heaviest leaf in the tree at time t.

5.1.1. Bounds on the Entropy-Based Criteria

Lemma 5. *The Shannon entropy function G_t^e at time t is bounded as $0 \leq G_t^e \leq (t+1)w \ln k$.*

Lemma 6. *The Gini-entropy function G_t^g at time t is bounded as $0 \leq G_t^g \leq (t+1)w\,(1-1/k)$.*

Lemma 7. *The modified Gini-entropy function G_t^m at time t is bounded as $\sqrt{\mathcal{C}-1} \leq G_t^m \leq (t+1)w\sqrt{k\mathcal{C}-1}$.*

The upper-bounds in Lemmas 5–7 are tight, where the equalities hold for the special case when $\forall_{i\in\{1,\dots,k\}, l\in\mathcal{L}_t}\pi_{l,i} = 1/k$, e.g., when each internal node of the tree produce a perfectly pure and balanced split.

5.1.2. Strong Concativity Properties of the Entropy-Based Criteria

So far we have been focusing on the time step t. Recall that n is the heaviest leaf at time t and its weight w_n is denoted by w for brevity. Consider splitting this leaf to two children n_0 and n_1. For ease of notation let $w_0 = w_{n_0}$ and $w_1 = w_{n_1}$, $\beta = P(h_n(x) > 0)$ and $P_i = P(h_n(x) > 0|i)$, and furthermore let π_i and h be the shorthands for $\pi_{n,i}$ and h_n, respectively. Recall that $\beta = \sum_{i=1}^{k} \pi_i P_i$ and $\sum_{i=1}^{k} \pi_i = 1$. Notice that $w_0 = w(1-\beta)$ and $w_1 = w\beta$. Let $\boldsymbol{\pi}$ be the k-element vector with i^{th} entry equal to π_i. Finally, let $\tilde{G}^e(\boldsymbol{\pi}) = \sum_{i=1}^{k} \pi_i \ln\left(\frac{1}{\pi_i}\right)$, $\tilde{G}^g(\boldsymbol{\pi}) = \sum_{i=1}^{k} \pi_i(1-\pi_i)$, and $\tilde{G}^m(\boldsymbol{\pi}) = \sum_{i=1}^{k} \sqrt{\pi_i(1-\pi_i)}$. Before the split the contribution of node n to resp. G_t^e, G_t^g, and G_t^m was resp. $w\tilde{G}^e(\boldsymbol{\pi})$, $w\tilde{G}^g(\boldsymbol{\pi})$, and $w\tilde{G}^m(\boldsymbol{\pi})$. Note that $\pi_{n_0,i} = \frac{\pi_i(1-P_i)}{1-\beta}$ and $\pi_{n_1,i} = \frac{\pi_i P_i}{\beta}$ are the probabilities that a randomly chosen x drawn from $\mathcal{P}$ has label i given that x reaches nodes n_0 and n_1 respectively. For brevity, let $\pi_{n_0,i}$ and $\pi_{n_1,i}$ be denoted respectively as $\pi_{0,i}$ and $\pi_{1,i}$. Let $\boldsymbol{\pi}_0$ be the k-element vector with i^{th} entry equal to $\pi_{0,i}$ and let $\boldsymbol{\pi}_1$ be the k-element vector with i^{th} entry equal to $\pi_{1,i}$. Notice that $\boldsymbol{\pi} = (1-\beta)\boldsymbol{\pi}_0 + \beta\boldsymbol{\pi}_1$. After the split the contribution of the same, now internal, node n changes to resp. $w((1-\beta)\tilde{G}^e(\boldsymbol{\pi}_0) + \beta\tilde{G}^e(\boldsymbol{\pi}_1))$, $w((1-\beta)\tilde{G}^g(\boldsymbol{\pi}_0) + \beta\tilde{G}^g(\boldsymbol{\pi}_1))$, and $w((1-\beta)\tilde{G}^m(\boldsymbol{\pi}_0) + \beta\tilde{G}^m(\boldsymbol{\pi}_1))$. We can compute the difference between the contribution of node n to the value of the entropy-based objectives in times t and $t+1$ as

$$\Delta_t^e = G_t^e - G_{t+1}^e = w\left[\tilde{G}^e(\boldsymbol{\pi}) - (1-\beta)\tilde{G}^e(\boldsymbol{\pi}_0) - \beta\tilde{G}^e(\boldsymbol{\pi}_1)\right], \tag{2}$$

$$\Delta_t^g = G_t^g - G_{t+1}^g = w\left[\tilde{G}^g(\boldsymbol{\pi}) - (1-\beta)\tilde{G}^g(\boldsymbol{\pi}_0) - \beta\tilde{G}^g(\boldsymbol{\pi}_1)\right], \tag{3}$$

$$\Delta_t^m = G_t^m - G_{t+1}^m = w\left[\tilde{G}^m(\boldsymbol{\pi}) - (1-\beta)\tilde{G}^m(\boldsymbol{\pi}_0) - \beta\tilde{G}^m(\boldsymbol{\pi}_1)\right]. \tag{4}$$

The next three lemmas, Lemmas 8–10, describe the strong concativity properties of the entropy, Gini-entropy and modified Gini-entropy, which can be used to lower-bound Δ_t^e, Δ_t^g, and Δ_t^m (Equations (2)–(4) correspond to a gap in the Jensen's inequality applied to the strongly concave function).

Lemma 8. *The Shannon entropy function $\tilde{G}^e$ is strongly concave with respect to l_1-norm with modulus 1, and thus the following holds $\tilde{G}^e(\boldsymbol{\pi}) - (1-\beta)\tilde{G}^e(\boldsymbol{\pi}_0) - \beta\tilde{G}^e(\boldsymbol{\pi}_1) \geq \frac{1}{2}\beta(1-\beta)\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}_1\|_1^2$.*

Lemma 9. *The Gini-entropy function $\tilde{G}^g$ is strongly concave with respect to l_2-norm with modulus 2, and thus the following holds $\tilde{G}^g(\boldsymbol{\pi}) - (1-\beta)\tilde{G}^g(\boldsymbol{\pi}_0) - \beta\tilde{G}^g(\boldsymbol{\pi}_1) \geq \beta(1-\beta)\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}_1\|_2^2$.*

Lemma 10. *The modified Gini-entropy function $\tilde{G}^m$ is strongly concave with respect to l_2-norm with modulus $\frac{2(\mathcal{C}-2)^2}{\mathcal{C}^3}$, and thus the following holds $\tilde{G}^m(\boldsymbol{\pi}) - (1-\beta)\tilde{G}^m(\boldsymbol{\pi}_0) - \beta\tilde{G}^m(\boldsymbol{\pi}_1) \geq \frac{(\mathcal{C}-2)^2}{\mathcal{C}^3}\beta(1-\beta)\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}_1\|_2^2$.*

Figure 3 illustrates different entropy criteria normalized to the $[0,1]$ interval.

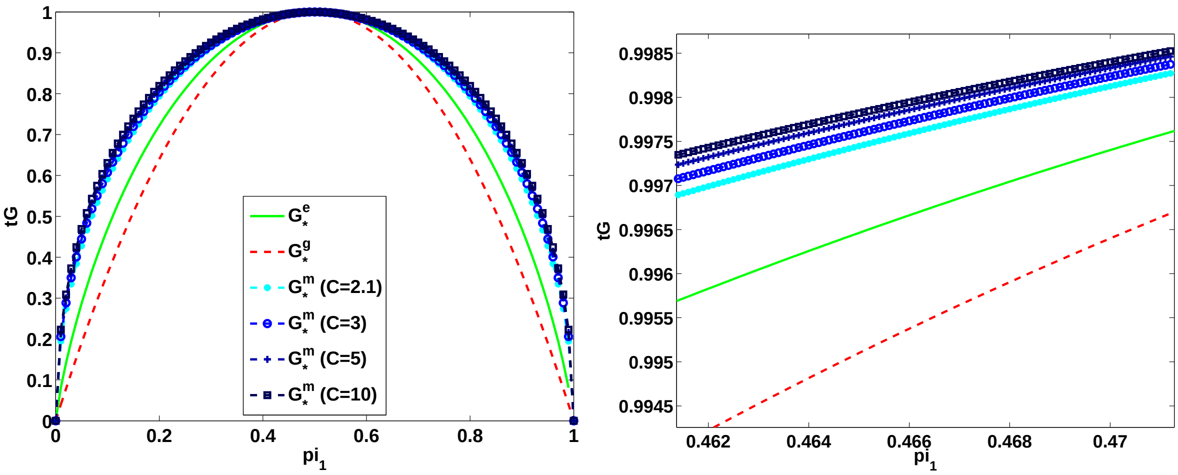

Figure 3. Functions $G^e_*(\pi_1) = \tilde{G}^e(\pi_1)/\ln 2 = \left(\pi_1 \ln\left(\frac{1}{\pi_1}\right) + (1-\pi_1)\ln\left(\frac{1}{1-\pi_1}\right)\right)/\ln 2$, $G^g_*(\pi_1) = 2\tilde{G}^g(\pi_1) = 4\pi_1(1-\pi_1)$, and $G^m_*(\pi_1) = (\tilde{G}^m(\pi_1) - \sqrt{\mathcal{C}-1})/(\sqrt{2*\mathcal{C}-1} - \sqrt{\mathcal{C}-1}) = (\sqrt{\pi_1(\mathcal{C}-\pi_1)} + \sqrt{(1-\pi_1)(\mathcal{C}-1+\pi_1)} - \sqrt{\mathcal{C}-1})/(\sqrt{2*\mathcal{C}-1} - \sqrt{\mathcal{C}-1})$ (functions $\tilde{G}^e(\pi_1)$, $\tilde{G}^g(\pi_1)$, and $\tilde{G}^m(\pi_1)$ were re-scaled to have values in $[0,1]$) as a function of π_1 (pi_1). Figure should be read in color.

5.2. Proof of Lemma 4 and Theorems 1–3

We finally proceed to proving all three boosting theorems, Theorems 1–3. Lemma 4 is a by-product of these proofs.

Proof. For the Shannon entropy it follows from Equation (2), Lemmas 5 and 8 that

$$\begin{aligned}
\Delta^e_t &\geq \frac{1}{2} w\beta(1-\beta)\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}_1\|_1^2 \\
&= \frac{1}{2}\frac{w}{\beta(1-\beta)}\left(\sum_{i=1}^{k} |\pi_i(P_i - \beta)|\right)^2 \\
&= \frac{wJ(h)^2}{8\beta(1-\beta)} \\
&\geq \frac{J(h)^2 G^e_t}{8\beta(1-\beta)(t+1)\ln k} \\
&\geq \frac{\gamma^2 G^e_t}{2(1-\gamma)^2(t+1)\ln k},
\end{aligned} \tag{5}$$

where the last inequality comes from the fact that $1-\gamma \geq \beta \geq \gamma$ (see the definition of γ in the weak hypothesis assumption) and $J(h) \geq 2\gamma$ (see weak hypothesis assumption). For the Gini-entropy criterion notice that from Equation (3), Lemmas 6, 9, and A4 it follows that

$$\begin{aligned}
\Delta^g_t &\geq w\beta(1-\beta)\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}_1\|_2^2 \\
&\geq \frac{1}{k} w\beta(1-\beta)\|\boldsymbol{\pi}_0 - \boldsymbol{\pi}_1\|_1^2 \\
&\geq \frac{\gamma^2 G^g_t}{(1-\gamma)^2(t+1)(k-1)},
\end{aligned} \tag{6}$$

where the last inequality is obtained similarly as the last inequality in Equation (5). And finally for the modified Gini-entropy it follows from Equation (4), Lemmas 7, 10, and A4 that

$$\begin{aligned}\Delta_t^m &\geq w\frac{(\mathcal{C}-2)^2}{\mathcal{C}^3}\beta(1-\beta)\|\boldsymbol{\pi}_0-\boldsymbol{\pi}_1\|_2^2\\ &\geq \frac{1}{k}w\frac{(\mathcal{C}-2)^2}{\mathcal{C}^3}\beta(1-\beta)\|\boldsymbol{\pi}_0-\boldsymbol{\pi}_1\|_1^2\\ &\geq \frac{\gamma^2 G_t^m}{\frac{\mathcal{C}^3}{(\mathcal{C}-2)^2}(1-\gamma)^2(t+1)k\sqrt{k\mathcal{C}-1}},\end{aligned} \tag{7}$$

where the last inequality is obtained as before.

Clearly the larger the objective $J(h)$ is at time t, the larger the entropy reduction ends up being. Let

$$\begin{aligned}\eta^e &= \frac{2\sqrt{2}\gamma}{(1-\gamma)\sqrt{\ln k}}, \quad \eta^g = \frac{4\gamma}{(1-\gamma)\sqrt{k-1}},\\ \eta^m &= \frac{4\gamma}{(1-\gamma)\sqrt{\frac{\mathcal{C}^3}{(\mathcal{C}-2)^2}k\sqrt{k\mathcal{C}-1}}}.\end{aligned} \tag{8}$$

For simplicity of notation assume Δ_t corresponds to either Δ_t^e, or Δ_t^g, or Δ_t^m, and G_t stands for G_t^e, or G_t^g, or G_t^m. Thus $\Delta_t > \frac{\eta^2 G_t}{16(t+1)}$, and we obtain

$$G_{t+1} \leq G_t - \Delta_t < G_t - \frac{\eta^2 G_t}{16(t+1)} = G_t\left(1-\frac{\eta^2}{16(t+1)}\right).$$

One can now compute the minimum number of splits required to reduce G_t below α, where $\alpha \in [0,1]$, from this recurrence inequality. Assume $\log_2(t+1) \in \mathbb{Z}^+$.

$$\begin{aligned}G_{t+1} &\leq G_t\left(1-\frac{\eta^2}{16(t+1)}\right)\\ &= G_1\left(1-\frac{\eta^2}{16\cdot 2}\right)\left(1-\frac{\eta^2}{16\cdot 3}\right)\dots(1-\frac{\eta^2}{16\cdot(t+1)})\\ &= G_1\left(1-\frac{\eta^2}{16\cdot 2}\right)\prod_{t'=3}^{4}\left(1-\frac{\eta^2}{16\cdot t'}\right)\dots\\ &\prod_{t'=(2^r/2)+1}^{2^r}\left(1-\frac{\eta^2}{16\cdot t'}\right)\dots\prod_{t'=(2^{\log_2(t+1)}/2)+1}^{2^{\log_2(t+1)}}\left(1-\frac{\eta^2}{16\cdot t'}\right),\end{aligned}$$

where $r = \{2,3,\dots,\log_2(t+1)\}$. Recall that

$$\begin{aligned}\prod_{t'=(2^r/2)+1}^{2^r}\left(1-\frac{\eta^2}{16\cdot t'}\right) &\leq \prod_{t'=(2^r/2)+1}^{2^r}\left(1-\frac{\eta^2}{16\cdot 2^r}\right)\\ &= \left(1-\frac{\eta^2}{16\cdot 2^r}\right)^{2^r/2} \leq e^{-\eta^2/32},\end{aligned}$$

where the last step follows from Lemma A5. Also note that by the same lemma $\left(1-\frac{\eta^2}{16\cdot 2}\right) \leq e^{-\eta^2/32}$. Thus,

$$G_{t+1} \leq G_1 e^{-\eta^2\log_2(t+1)/32}. \tag{9}$$

Therefore to reduce $G_{t+1} \leq \alpha$ (where α's are defined in Theorems 1–3) it suffices to make $t+1$ splits such that $\log_2(t+1) \geq \ln\left(\frac{G_1}{\alpha}\right)^{\frac{32}{\eta^2}}$ splits. Since $\log_2(t+1) = \ln(t+1) \cdot \log_2(e)$, where $e = \exp(1)$. Thus,

$$\ln(t+1) \geq \ln\left(\frac{G_1}{\alpha}\right)^{\frac{32}{\eta^2 \log_2(e)}} \Leftrightarrow t+1 \geq \left(\frac{G_1}{\alpha}\right)^{\frac{32}{\eta^2 \log_2(e)}}. \tag{10}$$

Recall that by resp. Lemmas 5–7 we have resp. $G_1^e \leq 2\ln k$, $G_1^g \leq 2(1-\frac{1}{k})$, $G_1^g \leq 2\sqrt{k\mathcal{C}-1}$. We consider the worst case setting (giving the largest possible number of split) thus we assume $G_1^e = 2\ln k$, $G_1^g = 2(1-\frac{1}{k})$, and $G_1^g \leq 2\sqrt{k\mathcal{C}-1}$. Combining that with Equations (8) and (10) yields statements of the main theorems. □

5.3. *Proof of Theorem 4*

We next proceed to directly proving the error bound. Recall that $\pi_{l,i}$ is the probability that the data point x corresponds to label i given that x reached l, i.e., $\pi_{l,i} = P(y(x) = i | x \text{ reached } l)$. Let the label assigned to the leaf be the majority label and thus lets assume that the leaf is assigned to label i if and only if the following is true $\forall_{\substack{z=\{1,2,\ldots,k\}\\ z\neq i}} \pi_{l,i} \geq \pi_{l,z}$. Therefore we can write that

$$\begin{aligned} \epsilon(\mathcal{T}) = \;& P(t(x) \neq y(x)) \\ = \;& \textstyle\sum_{l\in\mathcal{L}_t} w_l P(t(x) \neq y(x) | x \text{ reached } l) \end{aligned} \tag{11}$$

Let i_l be the majority label in leaf l, thus $\forall_{\substack{z=\{1,2,\ldots,k\}\\ z\neq i_l}} \pi_{l,i_l} \geq \pi_{l,z}$. We can continue as follows

$$\begin{aligned} \epsilon(\mathcal{T}) &= \sum_{l\in\mathcal{L}_t} w_l P(t(x) \neq i_l | x \text{ reached } l) \\ &= \sum_{l\in\mathcal{L}_t} w_l (1 - \pi_{l,i_l}) \\ &= \sum_{l\in\mathcal{L}_t} w_l (1 - \max(\pi_{l,1}, \pi_{l,2}, \ldots, \pi_{l,k}) \end{aligned} \tag{12}$$

Consider again the Shannon entropy $G(\mathcal{T})$ of the leaves of tree $\mathcal{T}$ that is defined as

$$\begin{aligned} G_t^e &= \sum_{l\in\mathcal{L}_t} w_l \sum_{i=1}^{k} \pi_{l,i} \ln \frac{1}{\pi_{l,i}} \\ G_t^e &= \frac{1}{log_2 e} \sum_{l\in\mathcal{L}_t} w_l \sum_{i=1}^{k} \pi_{l,i} \log_2 \frac{1}{\pi_{l,i}} \end{aligned}$$

Note that

$$\begin{aligned} G_t^e =& \frac{1}{log_2 e} \sum_{l\in\mathcal{L}_t} w_l \sum_{i=1} \pi_{l,i} \log_2 \frac{1}{\pi_{l,i}} \\ \geq& \frac{1}{log_2 e} \sum_{l\in\mathcal{L}_t} w_l \sum_{\substack{i=1\\ i\neq i_l}} \pi_{l,i} \log_2 \frac{1}{\pi_{l,i}} \\ \geq& \frac{1}{log_2 e} \sum_{l\in\mathcal{L}_t} w_l \sum_{\substack{i=1\\ i\neq i_l}} \pi_{l,i} \\ =& \frac{1}{log_2 e} \sum_{l\in\mathcal{L}_t} w_l (1 - \max(\pi_{l,1}, \pi_{l,2}, \ldots, \pi_{l,k})) \\ =& \frac{1}{log_2 e} \epsilon(\mathcal{T}), \end{aligned} \tag{13}$$

where the last inequality comes from the fact that $\forall_{\substack{i=\{1,2,\dots,\} \\ i\neq i_l}} \pi_{l,i} \leq 0.5$ and thus $\forall_{\substack{i=\{1,2,\dots,\} \\ i\neq i_l}} \frac{1}{\pi_{l,i}} \in [2;+\infty]$ and consequently $\forall_{\substack{i=\{1,2,\dots,\} \\ i\neq i_l}} \log_2 \frac{1}{\pi_{l,i}} \in [1;+\infty]$.

6. Conclusions

This paper aims at introducing theoretical tools, encapsulated in the boosting framework, that enable the comparison of different multi-class classification objective functions. The multi-class boosting is largely ununderstood from the theoretical perspective [5]. We provide an exhaustive theoretical analysis of the objective function underlying the recently proposed LOMtree algorithm for extreme multi-class classification and explore the connection of this objective to entropy-based criteria. We show that optimizing this objective simultaneously optimizes Shannon entropy, Gini-entropy and its modified variant, as well as the multi-class classification error. We expect that discussed tools can be used to obtain theoretical guarantees in the multi-label [28–30] and memory-constrained settings (we will explore this research direction in the future). We also consider extensions to different variants of the multi-class classification problem [31,32] and multi-output learning tasks [33,34]. We thus plan to build a unified theoretical framework for understanding extreme classification trees.

Author Contributions: A.C. derived the theoretical results and did the empirical evaluation. I.K.J. was working on improving the write-up of the paper and checking mathematical correctness.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Extreme Multiclass Classification Criteria

Appendix A.1. Numerical Experiments

We run the *LOMtree* algorithm, which is implemented in the open source learning system Vowpal Wabbit [35], on four benchmark multiclass data sets: *Mnist* (10 classes, downloaded from http://yann.lecun.com/exdb/mnist/), *Isolet* (26 classes, downloaded from http://www.cs.huji.ac.il/~shais/datasets/ClassificationDatasets.html), *Sector* (105 classes, downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html), and *Aloi* (1000 classes, downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html). The data sets were divided into training (90%) and testing (10%), where 10% of the training data set was used as a validation set. The regressors in the tree nodes are linear and were trained by SGD [36] with 20 epochs and the learning rate chosen from the set $\{0.25, 0.5, 0.75, 1, 2, 4, 8\}$. We investigated different swap resistances chosen from the set $\{4, 8, 16, 32, 64, 128, 256\}$. We selected the learning rate and the swap resistance as the one minimizing the validation error, where the number of splits in all experiments was set to 10 k.

Figure A1 shows the Shannon entropy, Gini-entropy, modified Gini-entropy (all normalized to the interval $[0, 1]$), and the multiclass classification error computed on the test data set as the function of the number of splits. The behavior of the Shannon entropy and Gini-entropy match the theoretical findings. However, the modified Gini-entropy instead drops the fastest with the number of splits, which in particular suggests that in this case perhaps tighter bounds could possibly be proved (for the binary case tighter analysis was shown in [25], but it is highly non-trivial to generalize this analysis to the multiclass classification setting). Furthermore, it can be observed that the behavior of the error closely mimics the behavior of the Gini-entropy. The Gini-entropy in all cases well-approximates the upper-bound on the error.

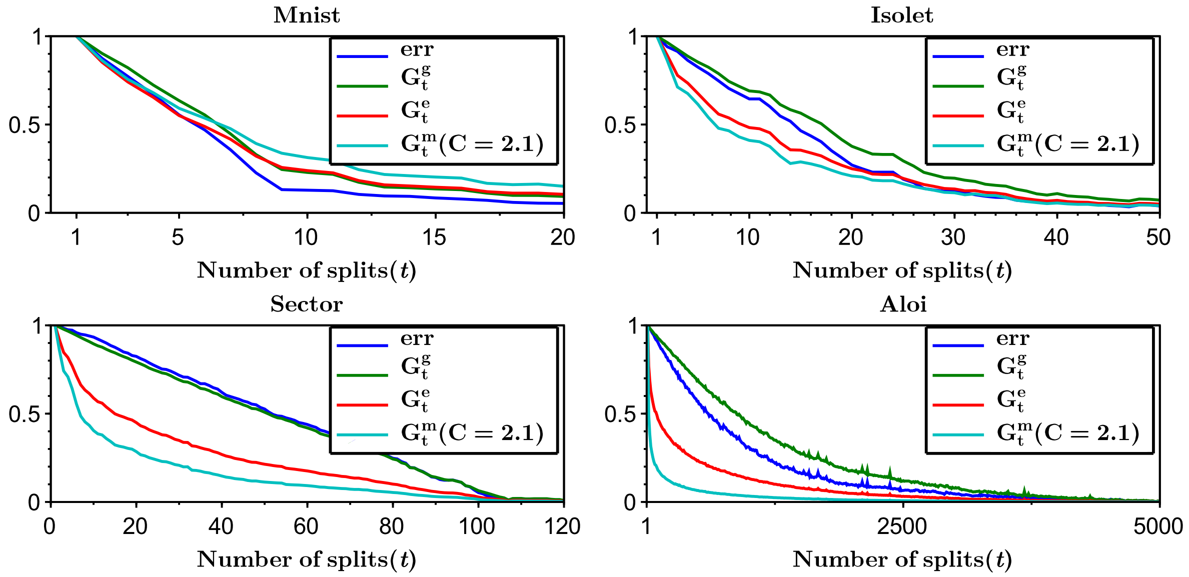

Figure A1. Functions G_t^e, G_t^g, and G_t^m, and the test error, all normalized to the interval $[0, 1]$, versus the number of splits. Figure is recommended to be read in color.

Appendix A.2. Additional Proofs

Proof of Lemma 1. The proof that if h induces a maximally pure and balanced partition then $J(h) = 1$ was done in [3] (Lemma 2) and is very basic. We focus here on the remaining part of statement, which is harder to show, and prove that if $J(h) = 1$ then h induces a maximally pure and balanced partition.

Without loss of generality assume each $\pi_i \in (0, 1)$. Recall that $\beta = P(h(x) > 0)$, and let $P_i = P(h(x) > 0|i)$. Also recall that $\beta = \sum_{i=1}^k \pi_i P_i$. Thus $J(h) = 2\sum_{i=1}^k \pi_i \left|\sum_{j=1}^k \pi_j P_j - P_i\right|$. The objective is certainly maximized in the extremes of the interval $[0, 1]$, where each P_i is either 0 or 1 (also note that at maximum, where $J(h) = 1$, it cannot be that all P_i's are 0 or all P_i's are 1). The function $J(h)$ is differentiable in these extremes ($J(h)$ is non-differentiable only when $\sum_{j=1}^k \pi_j P_j = P_i$, but at considered extremes the left-hand side of this equality is in $(0, 1)$, whereas the right-hand side is either 0 or 1). We then write

$$J(h) = 2\sum_{i\in\mathcal{P}} \pi_i \left(\sum_{j=1}^k \pi_j P_j - P_i\right) + 2\sum_{i\in\mathcal{N}} \pi_i \left(P_i - \sum_{j=1}^k \pi_j P_j\right),$$

where $\mathcal{P} = \{i : \sum_{j=1}^k \pi_j P_j \geq P_i\}$ and $\mathcal{N} = \{i : \sum_{j=1}^k \pi_j P_j < P_i\}$. Also let $\mathcal{P}^+ = \{i : \sum_{j=1}^k \pi_j P_j > P_i\}$ (clearly $\sum_{i\in\mathcal{P}^+} \pi_i \neq 1$ and $\sum_{i\in\mathcal{N}} \pi_i \neq 1$ in the extremes of the interval $[0, 1]$ where $J(h)$ is maximized). We then can compute the derivatives of $J(h)$ with respect to P_r, where $r = \{1, 2, \ldots, k\}$, everywhere where the function is differentiable as follows

$$\frac{\partial J}{\partial P_r} = \begin{cases} 2\pi_r(\sum_{i\in\mathcal{P}^+} \pi_i - 1) & \text{if } r \in \mathcal{P}^+ \\ 2\pi_r(1 - \sum_{i\in\mathcal{N}} \pi_i) & \text{if } r \in \mathcal{N} \end{cases},$$

and note that in the extremes of the interval $[0, 1]$ where $J(h)$ is maximized $\frac{\partial J}{\partial P_r} \neq 0$, since $\sum_{i\in\mathcal{P}^+} \pi_i \neq 1$, $\sum_{i\in\mathcal{N}} \pi_i \neq 1$, and each $\pi_i \in (0, 1)$. Since $J(h)$ is convex, and by the fact that in particular the derivative of $J(h)$ with respect to any P_r cannot be 0 in the extremes of the interval $[0, 1]$ where $J(h)$ is maximized, it follows that the $J(h)$ can only be maximized ($J(h) = 1$) at the extremes of the $[0, 1]$ interval. Thus we already proved that if $J(h) = 1$ then h induces a maximally pure partition. We are left with showing that if $J(h) = 1$ then h induces also a maximally balanced partition. We prove it by contradiction.

Assume $\beta \neq 0.5$. Denote as before $\mathcal{I}_0 = \{i : P(h(x) > 0|i) = 0\}$ and $\mathcal{I}_1 = \{i : P(h(x) > 0|i) = 1\}$. Recall $\beta = \sum_{i=1}^{k} \pi_i P_i = \sum_{i\in\mathcal{I}_0} \pi_i \cdot 0 + \sum_{i\in\mathcal{I}_1} \pi_i \cdot 1 = \sum_{i\in\mathcal{I}_1} \pi_i$. Thus,

$$\begin{aligned} J(h) = 1 &= 2\sum_{i\in\mathcal{I}_0} \pi_i\,|\beta| + 2\sum_{i\in\mathcal{I}_1} \pi_i\,|\beta - 1| \\ &= 2\beta\sum_{i\in\mathcal{I}_0} \pi_i + 2(1-\beta)\sum_{i\in\mathcal{I}_1} \pi_i \\ &= 2\beta(1 - \sum_{i\in\mathcal{I}_1} \pi_i) + 2(1-\beta)\sum_{i\in\mathcal{I}_1} \pi_i \\ &= 2\beta(1-\beta) + 2(1-\beta)\beta \\ &= -4\beta^2 + 4\beta < 1, \end{aligned}$$

where the last inequality comes from the fact that the quadratic form $-4\beta^2 + 4\beta$ is equal to 1 only when $\beta = 0.5$, and otherwise it is smaller than 1. Thus we obtain the contradiction which ends the proof. □

Proof of Lemma 2. We use the following notation: $\beta = P(h(x) > 0)$, and $P_i = P(h(x) > 0|i)$. Also let $\mathcal{P} = \{i : \beta \geq P_i\}$ and $\mathcal{N} = \{i : \beta < P_i\}$. Recall that $\beta = \sum_{i\in\{\mathcal{P}\cup\mathcal{N}\}} \pi_i P_i$, and $\sum_{i\in\{\mathcal{P}\cup\mathcal{N}\}} \pi_i = 1$. We split the proof into two cases.

- Let $\sum_{i\in\mathcal{P}} \pi_i \leq 1 - \beta$. Then

$$\begin{aligned} J(h) &= 2\sum_{i=1}^{k} \pi_i\,|\beta - P_i| \\ &= 2\sum_{i\in\mathcal{P}} \pi_i(\beta - P_i) + 2\sum_{i\in\mathcal{N}} \pi_i(P_i - \beta) \\ &= 2\sum_{i\in\mathcal{P}} \pi_i\beta - 2\sum_{i\in\mathcal{P}} \pi_i P_i + 2(\beta - \sum_{i\in\mathcal{P}} \pi_i P_i) \\ &\qquad - 2\beta(1 - \sum_{i\in\mathcal{P}} \pi_i) \\ &= 4\beta\sum_{i\in\mathcal{P}} \pi_i - 4\sum_{i\in\mathcal{P}} \pi_i P_i \\ &\leq 4\beta\sum_{i\in\mathcal{P}} \pi_i \leq 4\beta(1-\beta) \end{aligned}$$

  Thus $-4\beta^2 + 4\beta - J(h) \geq 0$ which, when solved, yields the lemma.
- Let $\sum_{i\in\mathcal{P}} \pi_i \geq 1 - \beta$ (thus $\sum_{i\in\mathcal{N}} \pi_i \leq \beta$). Note that $J(h)$ can be written as

$$J(h) = 2\sum_{i=1}^{k} \pi_i\,|P(h(x) \leq 0) - P(h(x) \leq 0|i)|\,,$$

  since $P(h(x) \leq 0) = 1 - P(h(x) > 0)$ and $P(h(x) \leq 0|i) = 1 - P(h(x) > 0|i)$. Let $\beta' = P(h(x) \leq 0) = 1 - \beta$, and $P_i' = P(h(x) \leq 0|i) = 1 - P_i$. Note that $\mathcal{P} = \{i : \beta \geq P_i\} = \{i : \beta' < P_i'\}$ and $\mathcal{N} = \{i : \beta < P_i\} = \{i : \beta' \geq P_i'\}$. Also note that $\beta' = \sum_{i\in\{\mathcal{P}\cup\mathcal{N}\}} \pi_i P_i'$. Thus

$$
\begin{aligned}
J(h) &= 2\sum_{i=1}^{k}\pi_i\left|\beta^{'}-P_i^{'}\right| \\
&= 2\sum_{i\in\mathcal{P}}\pi_i(P_i^{'}-\beta^{'})+2\sum_{i\in\mathcal{N}}\pi_i(\beta^{'}-P_i^{'}) \\
&= 2(\beta^{'}-\sum_{i\in\mathcal{N}}\pi_i P_i^{'})-2\beta^{'}(1-\sum_{i\in\mathcal{N}}\pi_i) \\
&\qquad\qquad +2\sum_{i\in\mathcal{N}}\pi_i\beta^{'}-2\sum_{i\in\mathcal{N}}\pi_i P_i^{'} \\
&= 4\beta^{'}\sum_{i\in\mathcal{N}}\pi_i-4\sum_{i\in\mathcal{N}}\pi_i P_i^{'}\leq 4\beta^{'}\sum_{i\in\mathcal{N}}\pi_i \\
&= 4(1-\beta)\sum_{i\in\mathcal{N}}\pi_i\leq 4\beta(1-\beta).
\end{aligned}
$$

Thus as before we obtain $-4\beta^2+4\beta-J(h)\geq 0$ which, when solved, yields the lemma.
□

Proof of Lemma 5. The lower-bound follows from the fact that the entropy of each leaf $\sum_{i=1}^{k}\pi_{l,i}\ln\left(\frac{1}{\pi_{l,i}}\right)$ is non-negative. We next prove the upper-bound.

$$
\begin{aligned}
G_t^e &= \sum_{l\in\mathcal{L}_t}w_l\sum_{i=1}^{k}\pi_{l,i}\ln\left(\frac{1}{\pi_{l,i}}\right) \\
&\leq \sum_{l\in\mathcal{L}_t}w_l\ln k\leq w\ln k\sum_{l\in\mathcal{L}_t}1 \\
&= (t+1)w\ln k,
\end{aligned}
$$

where the first inequality comes from the fact that uniform distribution maximizes the entropy, and the last equality comes from the fact that a tree with t internal nodes has $t+1$ leaves (also recall that w is the weight of the heaviest node in the tree at time t which is what we will also use in the next lemmas). □

Before proceeding to the actual proof of Lemma 6 we first introduce the helpful result captured in Lemma A1 and Corollary A1.

Lemma A1 (The inequality between Euclidean and arithmetic mean)**.** *Let $x_1,\ldots,x_k$ be a set of non-negative numbers. Then Euclidean mean upper-bounds the arithmetic mean as follows $\sqrt{\frac{\sum_{i=1}^{k}x_i^2}{k}}\geq\frac{\sum_{i=1}^{k}x_i}{k}$.*

Corollary A1. *Let $\{x_1,\ldots,x_k\}$ be non-negative. Then $\sum_{i=1}^{k}x_i^2\geq\frac{1}{k}\left(\sum_{i=1}^{k}x_i\right)^2$.*

Proof. By Lemma A1 we have $\sqrt{\frac{\sum_{i=1}^{k}x_i^2}{k}}\geq\frac{\sum_{i=1}^{k}x_i}{k}\Leftrightarrow\sum_{i=1}^{k}x_i^2\geq\frac{1}{k}\left(\sum_{i=1}^{k}x_i\right)^2$. □

Proof of Lemma 6. The lower-bound is straightforward since all $\pi_{l,i}$'s are non-negative. The upper-bound can be shown as follows (the last inequality results from Corollary A1):

$$\begin{aligned} G_t^g &= \sum_{l\in\mathcal{L}_t} w_l \sum_{i=1}^{k} \pi_{l,i}(1-\pi_{l,i}) \\ &\le w \sum_{l\in\mathcal{L}_t} \sum_{i=1}^{k} (\pi_{l,i} - \pi_{l,i}^2) = w \sum_{l\in\mathcal{L}_t} \left(1 - \sum_{i=1}^{k} \pi_{l,i}^2\right) \\ &\le w \sum_{l\in\mathcal{L}_t} \left(1 - \frac{1}{k}\left(\sum_{i=1}^{k} \pi_{l,i}\right)^2\right) = w \sum_{l\in\mathcal{L}_t} \left(1-\frac{1}{k}\right) \\ &= (t+1)w\left(1-\frac{1}{k}\right). \end{aligned}$$

□

Proof of Lemma 7. The lower-bound can be shown as follows. Recall that the function $\sum_{i=1}^{k}\sqrt{\pi_{l,i}(\mathcal{C}-\pi_{l,i})}$ is concave and therefore it is certainly minimized on the extremes of the $[0,1]$ interval, meaning where each $\pi_{l,i}$ is either 0 or 1. Let $I_0 = \{i : \pi_{l,i} = 0\}$ and let $I_1 = \{i : \pi_{l,i} = 1\}$. Thus $\sum_{i=1}^{k}\sqrt{\pi_{l,i}(\mathcal{C}-\pi_{l,i})} = \sum_{i\in I_1}\sqrt{\mathcal{C}-1} \ge \sqrt{\mathcal{C}-1}$. Combining this result with the fact that $\sum_{l\in\mathcal{L}_t} w_l = 1$ gives the lower-bound. We next prove the upper-bound. Recall that Lemma A1 implies that $(\sum_{i=1}^{k}\sqrt{\pi_{l,i}(\mathcal{C}-\pi_{l,i})})/k \le \sqrt{(\sum_{i=1}^{k}\pi_{l,i}(\mathcal{C}-\pi_{l,i}))/k}$, thus

$$\begin{aligned} G_t^m &= \sum_{l\in\mathcal{L}_t} w_l \sum_{i=1}^{k} \sqrt{\pi_{l,i}(\mathcal{C}-\pi_{l,i})} \\ &\le \sum_{l\in\mathcal{L}_t} w_l \sqrt{k\sum_{i=1}^{k}\pi_{l,i}(\mathcal{C}-\pi_{l,i})} \\ &= \sum_{l\in\mathcal{L}_t} w_l \sqrt{k\mathcal{C} - k^2\sum_{i=1}^{k}\frac{1}{k}\pi_{l,i}^2}. \end{aligned}$$

By Jensen's inequality $\sum_{i=1}^{k}\frac{1}{k}\pi_{l,i}^2 \ge (\sum_{i=1}^{k}\frac{1}{k}\pi_{l,i})^2 = \frac{1}{k^2}$. Thus

$$G_t^m \le \sum_{l\in\mathcal{L}_t} w_l\sqrt{k\mathcal{C}-1} \le (t+1)w\sqrt{k\mathcal{C}-1}.$$

□

Proof of Lemma 8. Lemma 8 is proven in [37] (Example 2.5). □

Lemma A2 (Lemma 14 in [38])**.** *If the function $\Phi(\boldsymbol{\pi})$ is twice differentiable, then the sufficient condition for strong concativity of Φ is that for all $\boldsymbol{\pi}$, $\boldsymbol{x}$, $\langle\nabla^2\Phi(\boldsymbol{\pi})\boldsymbol{x},\boldsymbol{x}\rangle \le -\sigma\|x\|^2$, where $\nabla^2\Phi(\boldsymbol{\pi})$ is the Hessian matrix of Φ at $\boldsymbol{\pi}$, and $\sigma > 0$ is the strong concativity modulus.*

Proof of Lemma 9. Note that $\langle\nabla^2\tilde{G}^g(\boldsymbol{\pi})\boldsymbol{x},\boldsymbol{x}\rangle \le -2\|\boldsymbol{x}\|_2^2$, and apply Lemma A2. □

Lemma A3 (Remark 2.2.4. in [39])**.** *The sum of strongly concave functions on $\mathbb{R}^n$ with modulus σ is strongly concave with the same modulus.*

Proof of Lemma 10. Consider functions $g(\pi_i) = \sqrt{f(\pi_i)}$, where $f(\pi_i) = \pi_i(\mathcal{C} - \pi_i)$, $\mathcal{C} \geq 2$, and $\pi_i \in [0,1]$. Also let $h(x) = \sqrt{x}$, where $x \in [0, \frac{\mathcal{C}^2}{4}]$. It is easy to see, using Lemma A2, that function f is strongly concave with respect to l_2-norm with modulus 2, thus

$$f(\theta\pi_i^{'} + (1-\theta)\pi_i^{''}) \geq \theta f(\pi_i^{'}) + (1-\theta)f(\pi_i^{''}) + \theta(1-\theta)\|\pi_i^{'} - \pi_i^{''}\|_2^2, \tag{A1}$$

where $\pi_i^{'}, \pi_i^{''} \in [0,1]$ and $\theta \in [0,1]$. Also note that h is strongly concave with modulus $\frac{2}{\mathcal{C}^3}$ in its domain $[0, \frac{\mathcal{C}^2}{4}]$ (the second derivative of h is $h^{''}(x) = -\frac{1}{4\sqrt{x^3}} \leq -\frac{2}{\mathcal{C}^3}$). The strong concativity of h implies that

$$\begin{aligned}\sqrt{\theta x_1 + (1-\theta)x_2} \geq \theta\sqrt{x_1} + (1-\theta)\sqrt{x_2}& \\ + \frac{1}{\mathcal{C}^3}\theta(1-\theta)\|x_1 - x_2\|_2^2,&\end{aligned}$$

where $x_1, x_2 \in [0, \frac{\mathcal{C}^2}{4}]$. Let $x_1 = f(\pi_i^{'})$ and $x_2 = f(\pi_i^{''})$. Then we obtain

$$\begin{aligned}\sqrt{\theta f(\pi_i^{'}) + (1-\theta)f(\pi_i^{''})} \geq \theta\sqrt{f(\pi_i^{'})} + (1-\theta)\sqrt{f(\pi_i^{''})}& \\ + \frac{1}{\mathcal{C}^3}\theta(1-\theta)\|f(\pi_i^{'}) - f(\pi_i^{''})\|_2^2.&\end{aligned} \tag{A2}$$

Note that

$$\begin{aligned}&\sqrt{f(\theta\pi_i^{'} + (1-\theta)\pi_i^{''})} \\ &\geq \sqrt{f(\theta\pi_i^{'} + (1-\theta)\pi_i^{''}) - \theta(1-\theta)\|\pi_i^{'} - \pi_i^{''}\|_2^2} \\ &\geq \sqrt{\theta f(\pi_i^{'}) + (1-\theta)f(\pi_i^{''})} \\ &\geq \theta\sqrt{f(\pi_i^{'})} + (1-\theta)\sqrt{f(\pi_i^{''})} \\ &\qquad + \frac{1}{\mathcal{C}^3}\theta(1-\theta)\|f(\pi_i^{'}) - f(\pi_i^{''})\|_2^2,\end{aligned}$$

where the second inequality results from Equation (A1) and the last (third) inequality results from Equation (A2). Finally note that the first derivative of f is $f^{'}(\pi_i) = \mathcal{C} - 2\pi_i \in [\mathcal{C} - 2, \mathcal{C}]$. Thus

$$\begin{aligned}&\frac{|f(\pi_i^{'}) - f(\pi_i^{''})|}{|\pi_i^{'} - \pi_i^{''}|} \geq \mathcal{C} - 2 \\ &\Leftrightarrow \|f(\pi_i^{'}) - f(\pi_i^{''})\|^2 \geq (\mathcal{C}-2)^2\|\pi_i^{'} - \pi_i^{''}\|^2,\end{aligned}$$

and combining this result with previous statement yields

$$\begin{aligned}&\sqrt{f(\theta\pi_i^{'} + (1-\theta)\pi_i^{''})} \\ &\geq \theta\sqrt{f(\pi_i^{'})} + (1-\theta)\sqrt{f(\pi_i^{''})} + \frac{(\mathcal{C}-2)^2}{\mathcal{C}^3}\theta(1-\theta)\|\pi_i^{'} - \pi_i^{''}\|^2,\end{aligned}$$

thus $g(\pi_i)$ is strongly concave with modulus $\frac{2(\mathcal{C}-2)^2}{\mathcal{C}^3}$. By Lemma A3, $\tilde{G}^m(\boldsymbol{\pi})$ is also strongly concave with the same modulus. □

The next two lemma are fundamental and they are used in the proof of Lemma 4 and the boosting theorems. The first one relates l_1-norm and l_2-norm and the second one is a simple property of the exponential function.

Lemma A4. *Let $x \in \mathbb{R}^k$ then $\|x\|_1 \leq \sqrt{k}\|x\|_2$.*

Lemma A5. *For $x \geq 1$ the following holds $\left(1 - \frac{1}{x}\right)^x \leq \frac{1}{e}$.*

References

1. Rifkin, R.; Klautau, A. In Defense of One-Vs-All Classification. *J. Mach. Learn. Res.* **2004**, *5*, 101–141.
2. Daume, H.; Karampatziakis, N.; Langford, J.; Mineiro, P. Logarithmic Time One-Against-Some. *arXiv* **2016**, arXiv:1606.04988.
3. Choromanska, A.; Langford, J. Logarithmic Time Online Multiclass prediction. In *Neural Information Processing Systems 2015*; Neural Information Processing Systems Foundation, Inc.: Vancouver, BC, Canada, 2015.
4. Schapire, R.E.; Freund, Y. *Boosting: Foundations and Algorithms*; The MIT Press: Cambridge, MA, USA, 2012.
5. Mukherjee, I.; Schapire, R.E. A theory of multiclass boosting. *J. Mach. Learn. Res.* **2013**, *14*, 437–497.
6. Beygelzimer, A.; Langford, J.; Ravikumar, P.D. Error-Correcting Tournaments. In *Algorithmic Learning Theory*; Springer: Berlin/Heidelberg, Germany, 2009.
7. Takimoto, E.; Maruoka, A. Top-down Decision Tree Learning As Information Based Boosting. *Theor. Comput. Sci.* **2003**, *292*, 447–464. [CrossRef]
8. Morin, F.; Bengio, Y. Hierarchical probabilistic neural network language model. *Aistats* **2005**, *5*, 246–252.
9. Bengio, S.; Weston, J.; Grangier, D. Label Embedding Trees for Large Multi-Class Tasks. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*; NIPS: Vancouver, BC, Canada, 2010.
10. Utgoff, P.E. Incremental Induction of Decision Trees. *Mach. Learn.* **1989**, *4*, 161–186. [CrossRef]
11. Domingos, P.; Hulten, G. *Mining High-speed Data Streams*; KDD: Boston, MA, USA, 2000.
12. Gama, J.; Rocha, R.; Medas, P. Accurate Decision Trees for Mining High-speed Data Streams. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–27 August 2003.
13. Beygelzimer, A.; Langford, J.; Lifshits, Y.; Sorkin, G.B.; Strehl, A.L. Conditional Probability Tree Estimation Analysis and Algorithms. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–21 June 2009.
14. Madzarov, G.; Gjorgjevikj, D.; Chorbev, I. A Multi-class SVM Classifier Utilizing Binary Decision Tree. *Informatica* **2009**, *33*, 225–233.
15. Weston, J.; Makadia, A.; Yee, H. Label Partitioning For Sublinear Ranking. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013.
16. Deng, J.; Satheesh, S.; Berg, A.C.; Fei-Fei, L. Fast and Balanced: Efficient Label Tree Learning for Large Scale Object Recognition. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*; NIPS: Vancouver, BC, Canada, 2011.
17. Zhao, B.; Xing, E.P. Sparse Output Coding for Large-Scale Visual Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013.
18. Hsu, D.; Kakade, S.; Langford, J.; Zhang, T. Multi-Label Prediction via Compressed Sensing. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*; NIPS: Vancouver, BC, Canada, 2009.
19. Agarwal, A.; Kakade, S.M.; Karampatziakis, N.; Song, L.; Valiant, G. Least Squares Revisited: Scalable Approaches for Multi-class Prediction. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014), Beijing, China, 21–26 June 2014.
20. Beijbom, O.; Saberian, M.; Kriegman, D.; Vasconcelos, N. Guess-Averse Loss Functions For Cost-Sensitive Multiclass Boosting. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014), Beijing, China, 21–26 June 2014.
21. Jernite, Y.; Choromanska, A.; Sontag, D. Simultaneous Learning of Trees and Representations for Extreme Classification and Density Estimation. *arXiv* **2017**, arXiv:1610.04658.
22. Mnih, A.; Hinton, G.E. A Scalable Hierarchical Distributed Language Model. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*; NIPS: Vancouver, BC, Canada, 2009.
23. Djuric, N.; Wu, H.; Radosavljevic, V.; Grbovic, M.; Bhamidipati, N. Hierarchical Neural Language Models for Joint Representation of Streaming Documents and their Content. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015.
24. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*; NIPS: Vancouver, BC, Canada, 2013.

25. Kearns, M.; Mansour, Y. On the Boosting Ability of Top-Down Decision Tree Learning Algorithms. In Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing (STOC '96), Philadelphia, PA, USA, 22–24 May 1996; reprinted in *J. Comput. Syst. Sci.* **1999**, *58*, 109–128. [CrossRef]
26. Breiman, L. *Classification Regression Trees*; Routledge: Abingdon, UK, 2017.
27. Quinlan, J.R. *C4.5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.
28. Liu, W.; Tsang, I.W. Making decision trees feasible in ultrahigh feature and label dimensions. *J. Mach. Learn. Res.* **2017**, *18*, 2814–2849.
29. Muñoz, E.; Nováček, V.; Vandenbussche, P.Y. Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models. *Brief. Bioinform.* **2017**, *20*, 190–202. [CrossRef] [PubMed]
30. Charte, F.; Rivera, A.J.; del Jesus, M.J.; Herrera, F. REMEDIAL-HwR: Tackling multilabel imbalance through label decoupling and data resampling hybridization. *Neurocomputing* **2019**, *326*, 110–122. [CrossRef]
31. Koster, C.H.; Seutter, M.; Beney, J. Multi-classification of patent applications with Winnow. In *International Andrei Ershov Memorial Conference on Perspectives of System Informatics*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 546–555.
32. Liu, W.; Tsang, I.W.; Müller, K.R. An easy-to-hard learning paradigm for multiple classes and multiple labels. *J. Mach. Learn. Res.* **2017**, *18*, 3300–3337.
33. Liu, W.; Xu, D.; Tsang, I.W.; Zhang, W. Metric learning for multi-output tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 408–422. [CrossRef] [PubMed]
34. Petersen, N.C.; Rodrigues, F.; Pereira, F.C. Multi-output bus travel time prediction with convolutional LSTM neural network. *Expert Syst. Appl.* **2019**, *120*, 426–435. [CrossRef]
35. Langford, J.; Li, L.; Strehl, A. Vowpal Wabbit (Fast Learning). 2007. Available online: http://hunch.net/~vw (accessed on 2 February 2019).
36. Bottou, L. Online Algorithms and Stochastic Approximations. In *Online Learning and Neural Networks*; Cambridge University Press: New York, NY, USA, 1998.
37. Shalev-Shwartz, S. Online Learning and Online Convex Optimization. *Found. Trends Mach. Learn.* **2012**, *4*, 107–194. [CrossRef]
38. Shalev-Shwartz, S. Online Learning: Theory, Algorithms, and Applications. Ph.D. Thesis, The Hebrew University of Jerusalem, Jerusalem, Israel, 2007.
39. Zhukovskiy, V. *Lyapunov Functions in Differential Games*; Stability and Control: Theory, Methods and Applications; Taylor & Francis: London, UK, 2003.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).