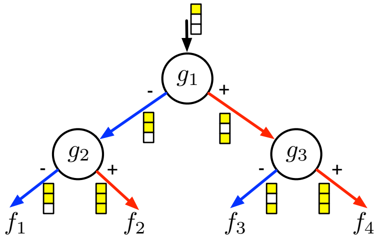
Supplementary Material: Model Selection by Linear Programming

Joseph Wang, Tolga Bolukbasi, Kirill Trapeznikov, and Venkatesh Saligrama

Boston University

1 Simple Tree Example

Here we present the derivation of the LP for a simple depth 2 tree below for the problem of supervised learning. Consider the decision system shown in Fig. 1. The goal is to learn the decision functions g_1, g_2, and g_3 that minimize the empirical risk.



$$\overset{\mathbf{P}}{\begin{bmatrix}0&0&0\\0&1&0\\1&0&0\\1&0&1\end{bmatrix}} - \overset{\mathbf{N}}{\begin{bmatrix}1&1&0\\1&0&0\\0&0&1\\0&0&0\end{bmatrix}} = \begin{bmatrix}-1&-1&0\\-1&+1&0\\+1&0&-1\\+1&0&+1\end{bmatrix}$$

Fig. 1: An example decision system of depth two: node $g_1(x_1)$ selects either to acquire sensor 2 for a cost c_2 or 3 for a cost c_3. Node $g_2(x_1, x_2)$ selects either to stop and classify with sensors $\{1, 2\}$ or to acquire 3 for c_3 and then stop. Node $g_3(x_1, x_3)$ selects to classify with $\{1, 3\}$ or with $\{1, 2, 3\}$.

In reformulating the risk, it is useful to define the "savings" for an example. The *savings*, π_k^i, for an example i, represents the difference between the worst case outcome, R_{max} and the risk $R_k(f_k, \mathbf{x}_i, y_i)$ for terminating and classifying at the kth leaf. The worst case risk is acquiring all sensors and incorrectly classifying: $R_{max} = 1 + \alpha \sum_m c_m$.

$$\pi_k^i = R_{max} - R_k(f_k, \mathbf{x}_i, y_i) = \mathbb{1}_{f_k(\mathbf{x}_i)=y_i} + \alpha \sum_{m \in S_k^C} c_m \tag{1}$$

Here, S_k^C is the complement set of sensors acquired along the path to leaf k (the sensors not acquired on the path to leaf k). Note that the savings do not depend on the decisions, $g_j' s$, that we are interested in learning.

For our example, there are only 4 leaf nodes and the state of terminating in a leaf is a encoded by a product of two indicators. For instance, to terminate in Leaf 1, $g_1(\mathbf{x}_i) \leq 0$ and $g_2(\mathbf{x}_i) \leq 0$. This empirical risk can be formulated by enumerating over the leaves and their associated risks:

$$
\begin{aligned}
R(\mathbf{g}, \mathbf{x}_i, y_i) = & \\
& \left.\Big(R_{max} - \pi_1^i\Big) \mathbb{1}_{g_1(\mathbf{x}_i) \leq 0} \mathbb{1}_{g_2(\mathbf{x}_i) \leq 0}\right\} \text{ Leaf 1} \\
& +\left.\Big(R_{max} - \pi_2^i\Big) \mathbb{1}_{g_1(\mathbf{x}_i) \leq 0} \mathbb{1}_{g_2(\mathbf{x}_i) > 0}\right\} \text{ Leaf 2} \\
& +\left.\Big(R_{max} - \pi_3^i\Big) \mathbb{1}_{g_1(\mathbf{x}_i) > 0} \mathbb{1}_{g_3(\mathbf{x}_i) \leq 0}\right\} \text{ Leaf 3} \\
& +\left.\Big(R_{max} - \pi_4^i\Big) \mathbb{1}_{g_1(\mathbf{x}_i) > 0} \mathbb{1}_{g_2(\mathbf{x}_i) > 0}\right\} \text{ Leaf 4}
\end{aligned} \tag{2}
$$

Directly replacing every $\mathbb{1}_{[z]}$ with an upper bounding surrogate such as a hinge loss, $\max[0, 1 + z] \geq \mathbb{1}_{[z]}$, produces a non-convex bilinear objective due the indicator product terms. Bilinear optimization is computationally intractable to solve globally.

Rather than directly substituting surrogates and solving the non-convex minimization problem, we reformulate the empirical risk with respect to the indicators in the following theorem:

Theorem 11 *The empirical risk in* (2) *is equal to* (3)*.*

$$
\begin{aligned}
R(g_1, g_2, g_3, \mathbf{x}_i, y_i) = & R_{max} - \pi_1^i - \pi_2^i - \pi_3^i - \pi_4^i + \\
\max \Big[& (\pi_3^i + \pi_4^i)\mathbb{1}_{g_1(\mathbf{x}_i) \leq 0} + \pi_2^i \mathbb{1}_{g_2(\mathbf{x}_i) \leq 0}, \\
& (\pi_3^i + \pi_4^i)\mathbb{1}_{g_1(\mathbf{x}_i) \leq 0} + \pi_1^i \mathbb{1}_{g_2(\mathbf{x}_i) > 0}, \\
& (\pi_1^i + \pi_2^i)\mathbb{1}_{g_1(\mathbf{x}_i) > 0} + \pi_4^i \mathbb{1}_{g_2(\mathbf{x}_i) \leq 0}, \\
& (\pi_1^i + \pi_2^i)\mathbb{1}_{g_1(\mathbf{x}_i) > 0} + \pi_3^i \mathbb{1}_{g_3(\mathbf{x}_i) > 0} \Big]
\end{aligned} \tag{3}
$$

Proof. Here, we provide a brief sketch of the proof. For full details please refer to Section 2. We utilize the following two identities: $\mathbb{1}_{[A]}\mathbb{1}_{[B]} = \min[\mathbb{1}_{[A]}, \mathbb{1}_{[B]}]$ and $\mathbb{1}_{[A]} = 1 - \mathbb{1}_{[\bar{A}]}$ and express the risk in (2) in terms of maximizations:

$$
\begin{aligned}
R\left(g_1, g_2, g_3, \mathbf{x}_i, y_i\right) = & R_{max} - \pi_1^i - \pi_2^i - \pi_3^i - \pi_4^i \\
& + \pi_1^i \max\left(\mathbb{1}_{g_1(\mathbf{x}_i) > 0}, \mathbb{1}_{g_2(\mathbf{x}_i) > 0}\right) \\
& + \pi_2^i \max\left(\mathbb{1}_{g_1(\mathbf{x}_i) > 0}, \mathbb{1}_{g_2(\mathbf{x}_i) \leq 0}\right) \\
& + \pi_3^i \max\left(\mathbb{1}_{g_1(\mathbf{x}_i) \leq 0}, \mathbb{1}_{g_3(\mathbf{x}_i) > 0}\right) \\
& + \pi_4^i \max\left(\mathbb{1}_{g_1(\mathbf{x}_i) \leq 0}, \mathbb{1}_{g_3(\mathbf{x}_i) \leq 0}\right)
\end{aligned} \tag{4}
$$

Recall that the signs of g_1, g_2, g_3 encode a unique path for $\mathbf{x}_i$. So let us consider sign patterns for each path. For instance, to reach leaf 1, $g_1 \leq 0$ and $g_2 \leq 0$. In this case, by inspection of (4), the risk is $(\pi_3^i + \pi_4^i)\mathbb{1}_{[g_1(\mathbf{x}_i) \leq 0]} + \pi_2^i \mathbb{1}_{[g_2(\mathbf{x}_i) \leq 0]} +$

constants. This is exactly the first term in the maximization in (3). We can perform such computation for each leaf (term in the max) in a similar fashion. And due to the interdependencies in (4), the term corresponding to a valid path encoding will be the maximizer in (3).

Risk Interpretability: Intuitively, in the reformulated empirical risk in (3), each term in the maximization encodes a path to one of the K leaves. The largest (active) term correspond to the path induced by the g_j's for an example $\mathbf{x}_i$. Additionally, the weights on the indicators in (3) represent the *savings lost* if the argument of the indicator is active. For example, if the decision function $g_1(\mathbf{x}_i)$ is negative, leaves 3 and 4 cannot be reached by $\mathbf{x}_i$, and therefore π_3^i and π_4^i, the savings associated with leaves 3 and 4, cannot be realized and are lost.

A distinct advantage of the reformulated risk in (3) arises when replacing indicators with convex upper-bounding surrogates of the form $\phi(z) \geq \mathbb{1}_{z \leq 0}$. Introducing such surrogates in the original risk in (2) produces a bilinear function for which a global optimum cannot be efficiently found. In contrast, introducing convex surrogate functions in (3) produces a convex upper-bound for the empirical risk.

2 Proof of Theorem 1

The product of indicators can be expressed as a minimization over the indicators, allowing the empirical loss to be expressed:

$$\begin{aligned}R\left(g_1, g_{21}, g_{22}, x_i, y_i\right) = \Bigg(1 + \sum_{k=1}^{K} c_k \\ - \pi_1^i \min\left(\mathbb{1}_{g_1(x_i) \leq 0}, \mathbb{1}_{g_{21}(x_i) \leq 0}\right) \\ - \pi_2^i \min\left(\mathbb{1}_{g_1(x_i) \leq 0}, \mathbb{1}_{g_{21}(x_i) > 0}\right) \\ - \pi_3^i \min\left(\mathbb{1}_{g_1(x_i) > 0}, \mathbb{1}_{g_{22}(x_i) \leq 0}\right) \\ - \pi_4^i \min\left(\mathbb{1}_{g_1(x_i) > 0}, \mathbb{1}_{g_{22}(x_i) > 0}\right) \Bigg).\end{aligned}$$

By swapping the inequalities in the arguments of the indicator functions, the minimization functions can be converted to maximization functions:

$$\begin{aligned}R\left(g_1, g_{21}, g_{22}, x_i, y_i\right) = \Bigg(1 + \sum_{k=1}^{K} c_k \\ + \pi_1^i \max\left(\mathbb{1}_{g_1(x_i) > 0}, \mathbb{1}_{g_{21}(x_i) > 0}\right) - \pi_1^i \\ + \pi_2^i \max\left(\mathbb{1}_{g_1(x_i) > 0}, \mathbb{1}_{g_{21}(x_i) \leq 0}\right) - \pi_2^i \\ + \pi_3^i \max\left(\mathbb{1}_{g_1(x_i) \leq 0}, \mathbb{1}_{g_{22}(x_i) > 0}\right) - \pi_3^i \\ + \pi_4^i \max\left(\mathbb{1}_{g_1(x_i) \leq 0}, \mathbb{1}_{g_{22}(x_i) \leq 0}\right) - \pi_4^i \Bigg).\end{aligned}$$

Note that due to the dependence of the indicators, there will always be 3 maximization terms equal to 1 and 1 maximization term equal to zero. As a result, the sum of maximizations can be expressed as a maximization over the 4 possible combinations, yielding the expression:

$$\begin{aligned}
&R\left(g_1, g_{21}, g_{22}, x_i, y_i\right) = \\
&\Bigg(1 + \sum_{k=1}^{K} c_k - \pi_1^i - \pi_2^i - \pi_3^i - \pi_4^i \\
&\max\Big((\pi_3^i + \pi_4^i)\mathbb{1}_{g_1(x_i)\leq 0} + \pi_2^i \mathbb{1}_{g_{21}(x_i)\leq 0}, \\
&(\pi_3^i + \pi_4^i)\mathbb{1}_{g_1(x_i)\leq 0} + \pi_1^i \mathbb{1}_{g_{21}(x_i)>0}, \\
&(\pi_1^i + \pi_2^i)\mathbb{1}_{g_1(x_i)>0} + \pi_4^i \mathbb{1}_{g_{21}(x_i)\leq 0}, \\
&(\pi_1^i + \pi_2^i)\mathbb{1}_{g_1(x_i)>0} + \pi_3^i \mathbb{1}_{g_{21}(x_i)>0}\Big)\Bigg).
\end{aligned}$$

3 Proof of Lemma 31

The product of indicators over an arbitrary binary tree is given by:

$$\begin{aligned}
&R(\mathbf{g}, \mathbf{x}_i, y_i) = \\
&\sum_{k=1}^{K} \overbrace{R_k(f_k, \mathbf{x}_i, y_i)}^{\text{risk of leaf } k} \underbrace{\prod_{j=1}^{K-1} \left[\mathbb{1}_{g_j(\mathbf{x}_i)>0}\right]^{\mathbf{P}_{k,j}} \left[\mathbb{1}_{g_j(\mathbf{x}_i)\leq 0}\right]^{\mathbf{N}_{k,j}}}_{\text{state of } G_k(\cdot) = \mathbf{x}_i \text{ in a tree}}.
\end{aligned}$$

Converting the product into a minimization over indicators, the function can be rewritten:

$$\begin{aligned}
&R(\mathbf{g}, \mathbf{x}_i, y_i) = \\
&\sum_{k=1}^{K} \left(R_{max} - \pi_k^i\right) \min_{j\in\{1,\ldots,K-1\}} \left(\left[\mathbb{1}_{g_j(\mathbf{x}_i)>0}\right]^{\mathbf{P}_{k,j}}, \left[\mathbb{1}_{g_j(\mathbf{x}_i)\leq 0}\right]^{\mathbf{N}_{k,j}}\right)
\end{aligned}$$

and using the identity $\mathbb{1}_A = 1 - \mathbb{1}_{\bar{A}}$, this can be converted to the maximization:

$$\begin{aligned}
&R(\mathbf{g}, \mathbf{x}_i, y_i) = R_{max} - \sum_{k=1}^{K} \pi_k^i + \\
&\sum_{k=1}^{K} \pi_k^i \max_{j\in\{1,\ldots,K-1\}} \left(\left[\mathbb{1}_{g_j(\mathbf{x}_i)\leq 0}\right]^{\mathbf{P}_{k,j}}, \left[\mathbb{1}_{g_j(\mathbf{x}_i)>0}\right]^{\mathbf{N}_{k,j}}\right).
\end{aligned}$$

As in the 2-region case, the dependence of the indicators always results in $K-1$ maximization terms equal to 1 and 1 maximization term equal to 0. By examination, the sum of maximization functions can be expressed as a single maximization over the paths of the leaves, resulting in a loss shown in (8).

4 Additional Explanation of Prop. 32

The linear program of Prop. 4.1 is constructed by replacing the indicators with hinge-losses of the appropriate signs:

$$\begin{aligned}
&\min_{\substack{g_1,\ldots,g_{K-1},\gamma^1,\ldots,\gamma^N\\ \alpha_1^1,\ldots,\alpha_{K-1}^N,\beta_1^1,\ldots,\beta_{K-1}^N}} \sum_{i=1}^{N} \gamma^i \text{ subject to:} \\
&\gamma^i \geq \mathbf{w}_{p,k}^i \begin{bmatrix} \alpha_1^i \\ \vdots \\ \alpha_{K-1}^i \end{bmatrix} + \mathbf{w}_{n,k}^i \begin{bmatrix} \beta_1^i \\ \vdots \\ \beta_{K-1}^i \end{bmatrix}, i \in [N],\ k \in [K] \\
&1 + g_j(\mathbf{x}_i) \leq \alpha_j^i,\ \ 1 - g_j(\mathbf{x}_i) \leq \beta_j^i, \quad \alpha_j^i \geq 0,\ \ \beta_j^i \geq 0, \\
&j \in [K-1], i \in [N]
\end{aligned} \tag{5}$$

Note that the linear program arises based on the fact that any maximization can be converted to a linear constraint with the introduction of a new variable. The maximization in the objective for each observation is replaced the first constraint, with the introduction of the variable γ^i. The maximization functions in the hinge losses are replaced by the second line of constraints, introducing the variables $\alpha_j^i = \max(1 + g_j(\mathbf{x}_i), 0)$ and $\beta_j^i = \max(1 - g_j(\mathbf{x}_i), 0)$.