Semi-Supervised Learning with Decision Trees: Graph Laplacian Tree Alternating Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Semi-supervised learning seeks to learn a machine learning model when only a small amount of the available data is labeled. The most widespread approach uses a graph prior, which encourages similar instances to have similar predictions. This has been very successful with models ranging from kernel machines to neural networks, but has remained inapplicable to decision trees, for which the optimization problem is much harder. We solve this based on a reformulation of the problem which requires iteratively solving two simpler problems: a supervised tree learning problem, which can be solved by the Tree Alternating Optimization algorithm; and a label smoothing problem, which can be solved through a sparse linear system. The algorithm is scalable and highly effective even with very few labeled instances, and makes it possible to learn accurate, interpretable models based on decision trees in such situations.

1 Introduction

Semi-supervised learning (SSL) is an important subfield of machine learning which has received a lot of attention in recent years given today's growing amount of data and widespread deployment of machine learning systems. One of the major reasons is that SSL is applicable when we have label scarcity issue. This is in contrast to the traditional fully supervised learning which requires access to the large amount of high-quality labeled data. However, obtaining such samples is often costly, time-consuming and sometimes even impractical. Therefore, SSL methods have received much praise in the machine learning literature [24] and they are widely used in many applications.

In this paper, we study the problem of training a decision tree (DT) model by leveraging a small percentage of labeled data and much larger sample of unlabeled data. A general strategy is to assume that similar instances have similar predictions which is commonly incorporated into an objective as a graph prior (e.g. graph Laplacian). *Why trees?* First, trees are considered to be interpretable models since the prediction is obtained by routing an input to the unique root-to-leaf path which can be reformulated as if-then rules. Second, they are widely used in wide spectrum of applications, such as data mining [18], computer vision [6], finance, etc. They are typically employed as base learners in an ensemble (e.g. bagging [2] or boosting [8]). However, they are used as standalone predictors as well. For instance, recent algorithmic advances in training non-greedy trees have shown that oblique trees can perform competitively well in number of tasks and strike a good balance between accuracy and interpretability [4].

However, similar to many non-linear methods, DTs are well-known to overfit for small sized (labeled) data which is the case in SSL. As an illustration, consider fig. 1 that shows artificially generated binary classification problem in 2D. An oblique tree achieves a certain good performance when it was provided the entire population of labeled data. But the error significantly increases if a tree is trained on six labeled instances only (center-right plot). Whereas the benefit is self-evident when

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

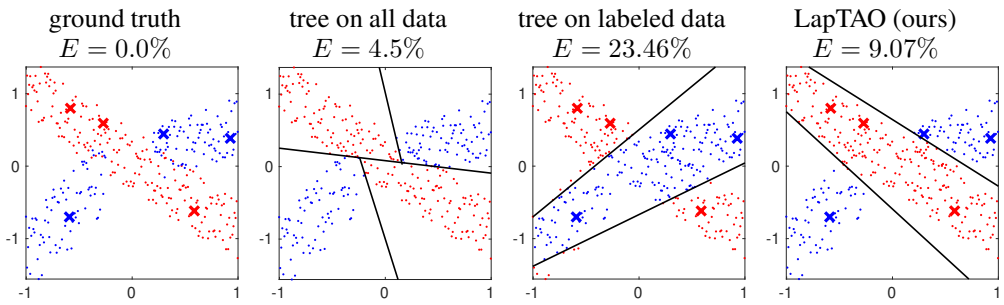

Figure 1: Binary classification on 2D. The leftmost plot shows the original data and corresponding class labels. Crossed markers (six in total) indicate the labeled points that we provide to any given SSL algorithm. "tree on all data" shows the decision boundary obtained by using all available labeled data, whereas center-right plot uses only six labeled points to train a tree. The rightmost plot shows the result of our SSL framework. All trees are *oblique* of depth $\Delta = 2$.

we provide all data (six labeled and the rest are unlabeled) and properly optimize a tree within SSL framework. In our proposed approach, we first state the objective which consists of supervised loss (for labeled data only) and the graph Laplacian regularization (also known as manifold regularization [1]). The resulting optimization problem is long considered to be hard to solve since trees define non-differentiable and non-convex mapping. By reformulating the problem as constrained optimization, we derive an efficient and scalable iterative algorithm (section 3) which requires solving two simpler problems at each step: a sparse linear system and a supervised tree learning problem. For the latter part, we use Tree Alternating Optimization algorithm [4] which is crucial for the success of our approach (section 3.1). Moreover, for a special case where the tree structure as well as parameters in each decision node (not leaves) are fixed, we derive the exact solution given by another linear system (section 3.3). Experimental results (section 4) suggest that the algorithm is able to learn accurate and interpretable decision trees even with very few labeled instances.

2 Related work

Although the literature on this topic is immense (see [24, 17]), incorporating decision trees into SSL has received a little attention. One possible explanation could be intractability of the optimization problem: majority of methods in SSL are based on adding a graph prior (or similarity matrix) as a regularization which exploits the geometry of underlying data distribution [1, 23, 25, 26], and minimizing an objective with such regularization is non-trivial task when we have decision trees as a predictive model.

That said, several attempts have been made to apply SSL for trees. Levatić et al. [12] modify a splitting criterion for recursive tree induction by taking into account unlabeled data. That is, a splitting score for each [feature, threshold] pair consists of two parts: traditional impurity score (e.g. gini index) for labeled data and a "clustering" score for all available data which pushes each child to have instances of the same cluster. Tanha et al. [16] apply self-training framework [21] to train classification trees with labeled and unlabeled data. Note that the self-training is a generic framework and can be applied with any classifier which produces the class prediction probabilities. It trains a classifier iteratively where the first iteration includes only labeled data. After that, it uses model predictions and adds the most "confidently" predicted instances to enrich the amount of supervision. Iteration continues by increasing the quantity of such "labeled" data at each step. Finally, Kemp et al. [11] propose Bayesian approach where unlabeled data assist in inferring a latent tree structure from distribution over trees. The limitations of this approach include: 1) it can be applied to classification tasks only; 2) it uses Markov Chain Monte Carlo (MCMC) techniques to sample from distribution over trees which leads to noticeable scalability issues on large data.

Unlike all these methods, we attempt to directly minimize a regularized SSL objective (with a graph prior) over the parameters of a decision tree. The method is applicable to both regression and classification problems, and can train trees of arbitrary complexity (i.e., axis-aligned, oblique, etc.). Next, we describe our approach in detail.

3 LapTAO: semi-supervised learning framework for decision trees

We are given the dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$, where $\mathcal{D}_l = \{\mathbf{x}_n, y_n\}_{n=1}^{l}$ is the labeled portion of data with size l and $\mathcal{D}_u = \{\mathbf{x}_n\}_{n=l+1}^{N}$ is unlabeled portion of data with size $N - l$. Then our goal is to minimize the following regularized objective:

$$E(\mathbf{\Theta}) = \frac{1}{l}\sum_{n=1}^{l}(T(\mathbf{x}_n;\mathbf{\Theta}) - y_n)^2 + \alpha\,\phi(\mathbf{\Theta}) + \frac{\gamma}{N^2}\sum_{n,m=1}^{N} w_{nm}\left(T(\mathbf{x}_n;\mathbf{\Theta}) - T(\mathbf{x}_m;\mathbf{\Theta})\right)^2. \quad (1)$$

Here, w_{nm} are weights in the affinity matrix (or similarity matrix), $T\colon \mathbb{R}^D \to \mathbb{R}$ is the tree predictive mapping with parameters $\mathbf{\Theta} = \{\boldsymbol{\theta}_i\}_{\text{nodes}}$. Next, $\phi(\cdot)$ is the regularization penalty such as $\|\cdot\|_1$ and γ,α are regularization hyperparameters. We emphasize that we consider T as a parametric model with trainable weights in each node (like fixing a neural net architecture and optimizing over its parameters). If T was differentiable, one could optimize (1) via gradient-based methods as it was done for neural nets [19]. Similarly, the solution is relatively straightforward to obtain if the overall problem in (1) is convex [1]. However, this is non-trivial with a tree which defines a non-differentiable and non-convex mapping. Instead, we will reformulate the problem as follows. We introduce a new variable z for each training instance and consider the constrained problem:

$$\min_{z_1,\ldots,z_N,\mathbf{\Theta}} \frac{1}{l}\sum_{n=1}^{l}(z_n - y_n)^2 + \alpha\,\phi(\mathbf{\Theta}) + \frac{\gamma}{N^2}\sum_{n,m=1}^{N} w_{nm}(z_n - z_m)^2 \quad (2)$$

$$\text{s.t.} \quad z_n = T(\mathbf{x}_n;\mathbf{\Theta}) \quad n = 1,\ldots,N. \quad (3)$$

Obviously, by putting constraints (3) back to the eq. (2), we end up with the same objective as in (1), so these two problems are equivalent. Let us denote $\mathbf{y} = [y_1, y_2, \ldots, y_l, 0, 0, \ldots]^T \in \mathbb{R}^N$, as the augmented ground truth vector, i.e., we simply put zeros for the unlabeled fraction of data. Similarly, introduce a diagonal matrix $\mathbf{J} = \text{diag}(1,\ldots,1,0,\ldots,0) \in \mathbb{R}^{N\times N}$ with the first l diagonal entries equal to 1 and the rest 0. Also, introduce graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$ where diagonal matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ with entries $d_{nn} = \sum_{m=1}^{N} w_{nm}$ and $\mathbf{W} = (w_{nm}) \in \mathbb{R}^{N\times N}$ is the affinity matrix. Finally, denote $\mathbf{z} = [z_1,\ldots,z_N]^T$ and $\mathbf{t}(\mathbf{X};\mathbf{\Theta}) = [T(\mathbf{x}_1;\mathbf{\Theta}),\ldots,T(\mathbf{x}_N;\mathbf{\Theta})]^T$ where $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_N)$. Then we rewrite eq. (2)-(3) as follows:

$$\min_{\mathbf{z},\mathbf{\Theta}} (\mathbf{z}-\mathbf{y})^T\mathbf{J}\,(\mathbf{z}-\mathbf{y}) + \alpha\,\phi(\mathbf{\Theta}) + \gamma\,\mathbf{z}^T\mathbf{L}\,\mathbf{z} \quad \text{s.t.} \quad \mathbf{z} = \mathbf{t}(\mathbf{X};\mathbf{\Theta}). \quad (4)$$

Here, we absorb some constants (e.g. normalization $\frac{1}{l}$) into matrices. We solve this using the *augmented Lagrangian* [14] method which defines a new, unconstrained optimization problem:

$$\min_{\mathbf{z},\mathbf{\Theta}} (\mathbf{z}-\mathbf{y})^T\mathbf{J}\,(\mathbf{z}-\mathbf{y}) + \alpha\,\phi(\mathbf{\Theta}) + \gamma\,\mathbf{z}^T\mathbf{L}\,\mathbf{z} - \boldsymbol{\lambda}^T(\mathbf{z} - \mathbf{t}(\mathbf{X};\mathbf{\Theta})) + \mu\|\mathbf{z} - \mathbf{t}(\mathbf{X};\mathbf{\Theta})\|^2 \quad (5)$$

where $\boldsymbol{\lambda} \in \mathbb{R}^N$ are the estimates of Lagrange multipliers. Optimizing this for fixed $\mu > 0$ produces the sequence of $(\mathbf{z}_\mu, \mathbf{t}_\mu(\mathbf{X};\mathbf{\Theta}))$ and as $\mu \to \infty$, we force the minimizer to be in the feasible region for the constrained problem. Finally, we apply alternating optimization to minimize (5) over $\mathbf{z}$ and $\mathbf{t}(\mathbf{X};\mathbf{\Theta})$:

- **Label–step** (optimizing over $\mathbf{z}$ given fixed $\mathbf{t}(\mathbf{X};\mathbf{\Theta})$). The objective in eq. (5) is a quadratic function and a minimizer is obtained by solving the linear system:

$$\begin{aligned}&\min_{\mathbf{z}} (\mathbf{z}-\mathbf{y})^T\mathbf{J}\,(\mathbf{z}-\mathbf{y}) + \gamma\,\mathbf{z}^T\mathbf{L}\,\mathbf{z} - \boldsymbol{\lambda}^T(\mathbf{z} - \mathbf{t}(\mathbf{X};\mathbf{\Theta})) + \mu\|\mathbf{z} - \mathbf{t}(\mathbf{X};\mathbf{\Theta})\|^2 \Rightarrow \\ &\mathbf{A}\mathbf{z} = \mathbf{J}\mathbf{y} + \mu\mathbf{t}(\mathbf{X};\mathbf{\Theta}) + \frac{1}{2}\boldsymbol{\lambda}\end{aligned} \quad (6)$$

  where $\mathbf{A} = \mathbf{J} + \mu\mathbf{I} + \gamma\mathbf{L}$ is a positive definite matrix (see the supplementary materials). Moreover, it is easy to see that $\mathbf{A}$ will be a sparse matrix if graph Laplacian $\mathbf{L}$ is sparse. And this is the case since in practice we construct $\mathbf{W}$ by using the nearest neighbors graph. This allows us to solve the large scale linear system in an efficient way (e.g. via Conjugate Gradient method). The most costly part will be the construction of $\mathbf{W}$. However, it only depends on input instances (without $\mathbf{y}$) and it is computed only once before we start the alternating optimization. Intuitively, the "label–step" can be interpreted as "approximating" the labels (for $\mathcal{D}_u$) using the graph Laplacian and predictions obtained from the current tree (i.e., label smoothing).

- **Tree–step** (optimizing over $\mathbf{\Theta}$ given fixed $\mathbf{z}$). The problem (5) reduces to a regression fit of a tree:

$$\begin{aligned} &\min_{\mathbf{\Theta}} \ \mu\|\mathbf{z} - \mathbf{t}(\mathbf{X};\mathbf{\Theta})\|^2 + \alpha\,\phi(\mathbf{\Theta}) - \boldsymbol{\lambda}^T(\mathbf{z} - \mathbf{t}(\mathbf{X};\mathbf{\Theta})) \Leftrightarrow \\ &\min_{\mathbf{\Theta}} \left\| \left(\mathbf{z} - \frac{1}{2\mu}\boldsymbol{\lambda}\right) - \mathbf{t}(\mathbf{X};\mathbf{\Theta}) \right\|^2 + \frac{\alpha}{\mu}\phi(\mathbf{\Theta}). \end{aligned} \tag{7}$$

  Note that here we use $(\mathbf{z} - \frac{1}{2\mu}\boldsymbol{\lambda})$ as labels (not y_n which is not defined for $\mathcal{D}_u$ anyway). We solve this problem using Tree Alternating Optimization (TAO) algorithm which we describe in section 3.1 below. Intuitively, this step can be understood as fitting a tree with the current estimates of the labels.

Finally, the step over Lagrange multipliers is done by $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda} - \mu(\mathbf{z} - \mathbf{t}(\mathbf{X};\mathbf{\Theta}))$. In summary, our algorithm alternates between solving a linear system and training a tree. After each (label,tree)–step, the penalty term μ increases, $\boldsymbol{\lambda}$ is updated and we keep iterating until convergence or other stopping criterion (e.g. maximum number of iterations). We denote our proposed iterative algorithm as *LapTAO* and provide detailed pseudocode in the supplementary materials.

Initialization for LapTAO To start our iterative algorithm, we need to obtain initial solutions $(\mathbf{z}_0,\ \mathbf{t}_0)$ for (5) when $\mu \to 0^+$. This is straightforward to achieve for $\mathbf{z}_0$ as it involves solving the same linear system as in (6) but with $\mu = 0$ and $\boldsymbol{\lambda} = \mathbf{0}$: $(\mathbf{J} + \gamma\mathbf{L})\mathbf{z} = \mathbf{J}\mathbf{y}$. This can be considered as non-parametric smoothing of the labels by diffusing them by graph Laplacian. After that, we proceed with fitting a tree ("tree–step") using $\mathbf{z}_0$. Note that TAO requires an initial tree (see section 3.1) which we obtain by generating a complete tree of depth Δ with random vector (sampled from normal distribution) at each node.

Extension to classification and multioutput regression We can extend LapTAO for multiple outputs in a straightforward way: the "label–step" solves the linear system (6) for each dimension separately and the "tree–step" applies TAO in a usual way as it can handle a vector valued output. As for the classification task, we use one-hot encoding of the labels and consider the problem as a regression task which is commonly done in decision tree literature [9] (especially for boosted trees). It is possible to extend our framework for other losses (e.g. hinge, logistic) which may work better for classification task. This requires certain changes in the "label–step" and we will explore it in our future works.

Extension to other models The way LapTAO works is simple: solve a linear system to get the "smoothed" labels and fit a regressor on them, repeat these two steps while increasing μ. Although we derive our algorithm specifically for decision trees, it is obvious that it can be extended to other machine learning models: random forests, gradient boosted trees, neural networks, etc. The only part that changes is the "tree–step" which will then involve fitting another model on smoothed labels.

3.1 Overview of the tree alternating optimization (TAO) algorithm

Potentially, one could apply any tree fitting algorithm to solve the "tree–step" in LapTAO. Examples include: CART [3], C5.0 [15], OC1 [13], etc. But there is one important caveat. It is known that alternating optimization is the most effective when the step over each block is (ideally) exact. This is achievable for the "label–step" since we can handle the linear system. However, training a tree optimally even in the simplest case (axes-aligned with binary inputs and output) is NP-hard [10]. Therefore, one can rely on a "good" approximate solution. This is hard to achieve with greedy tree induction procedure (e.g. CART) due to suboptimality [9]. In contrast, the TAO algorithm fits decision trees in non-greedy fashion and generally finds better optima of tree learning problem. It scales well to large data and can train trees of arbitrary complexity (e.g. oblique, axis-aligned, neural). Another important advantage of TAO is that it starts from pre-determined initial tree allowing a warm-start, whereas there is no notion of warm start for CART-style methods. This allows to efficiently leverage the fitted tree from previous iteration as we increase μ.

Below we provide a summary of TAO for the sake of completing the picture of our proposed approach. More details can be found in [4]. TAO considers decision tree as a parametric model with the given structure and trainable weights in each node (e.g. $\mathbf{\Theta} = \{\boldsymbol{\theta}_i\}_{\text{nodes}}$) and optimizes an objective as in (7)). It is important to note that TAO requires the loss to be the sum over *individual*

points, so it cannot handle (1) which includes pairwise distances. However, our reformulation of the problem allows us to apply this framework since the minimization in 7 is now in desirable form.

TAO takes as input an initial tree which can be obtained from random initialization (e.g. a complete tree of depth Δ with random weights at each node) or from other approximate algorithm. Next, it uses a *separability condition* to decompose the loss function (7) over the subsets of non-descendant nodes. For instance, all nodes at the same depth are non-descendants w.r.t. each others and can be trained independently (and in parallel). Following this logic, TAO uses the following alternating optimization scheme: 1) pick all nodes at a certain depth and fix all remaining nodes (i.e., we consider them as a function with constant parameters); 2) optimize each node at that depth efficiently by solving the *reduced problem* (see below); 3) repeat this for all depths. One pass through all nodes is called a *TAO iteration* and we keep iterating until convergence occurs or until we reach the maximum number of iterations (pseudocode is given in the supplementary materials). The notion of reduced problem comes from the fact that we optimize a single node i over its weights given a subset of training points which currently reach i (denoted as reduced set $\mathcal{R}_i$). The solution of the problem depends on its type:

- **Decision node**. Since we focus on binary trees, this always involves solving binary classification problem. Given the reduced set $\mathcal{R}_i$, we pass each input $\{\mathbf{x}_n, y_n\} \in \mathcal{R}_i$ to both children and calculate the error $(e_n^{\text{left}}, e_n^{\text{right}})$ induced by each child (note that all descendant nodes and their parameters are fixed). Next, assign a pseudolabel $\bar{y}_n \in \{-1, 1\}$ depending on which child brings the lower error (either left or right). Each pseudolabel comes with the weight $|e_n^{\text{left}} - e_n^{\text{right}}|$ since the error made for each instance is different. Solve the weighted binary classification problem on $\{\mathbf{x}_n, \bar{y}_n\}$.
- **Leaf**. Note that the actual prediction of a tree is given by its leaves and decision nodes are just responsible for routing a point to the corresponding leaf. Therefore, optimization problem for a leaf is equivalent to the original problem (7) but on its reduced set. A specific solution depends on what type of model we use. If leaves are constant, then we take the average of response values (y) in that leaf. If we use linear leaves, then the solution is fitting a linear regressor on $\mathcal{R}_i$ (possibly with regularization).

3.2 Sparse oblique decision trees

Although the TAO algorithm is applicable to a large spectrum of decision tree types, we pick a *sparse oblique tree with constant leaves* as our main model. Here, each decision node i makes a hyperplane-based split: go to left child if $\mathbf{w}^T\mathbf{x} < w_0$, else go to right. Therefore, the decision node optimization from section 3.1 involves fitting a linear classifier via logistic regression. Additionally, we apply ℓ_1 penalty as a regularization term $\phi(\cdot)$ to encourage sparsity. We use LIBLINEAR [7] to fit this model. As for the leaves, we take the average of response values (y) in that leaf.

The motivation behind choosing sparse oblique trees is twofold. First, traditional axis-aligned trees are restrictive since each split uses a single feature neglecting interactions or correlations between features. Indeed, empirical results suggest that oblique trees are more powerful and generally achieve better performance [22]. Besides, thanks to the sparsity, only few features are active at each split. This, together with the fact that oblique trees typically have small depth, makes the final model more interpretable.

Computational complexity for LapTAO It can be clearly seen that the inference time for our model (sparse oblique decision tree) is lightning fast: since each input follows only one root-to-leaf path, it is proportional to the depth times sparse vector multiplication at each node. As for the training time, we need to consider two components. For solving large (sparse) linear system, we use Conjugate Gradient method which iterates at most N times (usually much smaller than that and can be improved via preconditioning) and each iteration takes $O(N^2)$. However, this is significantly cheaper than $O(N^2)$ with sparse matrices. In practice, this step takes at most 30 seconds on the largest experiment we conducted. The most costly part of the algorithm is fitting a tree with TAO which makes several passes through all nodes of a tree. Its complexity is upper bounded by the tree depth times the cost of solving the logistic regression on the whole training set [4]. So, it is still manageable runtime and can be parallelized allowing to scale on large data.

3.3 Special case: exact solution when the tree structure is fixed

As we have mention earlier, the problem (1) is hard to solve over the entire space of decision trees. However, it is possible to obtain an exact solution for some special cases. Particularly, if we fix the tree structure as well as parameters in each decision node, then the problem reduces to solving another linear system. In this case, the only parameters to optimize are located in leaves (assuming each leaf outputs a constant value). We can reformulate the tree prediction as the sum of basis functions: $T(\mathbf{x}) = \sum_{i=1}^{m} c_i \, b_i(\mathbf{x})$, where m is the number of leaves and $b_i(\cdot) \in \{0, 1\}$ is 1 only if $\mathbf{x}$ ends up in leaf i. Now we can rewrite (1) as the following minimization problem over parameters of all leaves $\mathbf{c} = (c_1, \ldots, c_m)$:

$$E(\mathbf{c}) = \frac{1}{l} \sum_{n=1}^{l} \left(\sum_{i=1}^{m} c_i \, b_i(\mathbf{x}_n) - y_n \right)^2 + \frac{\gamma}{N^2} \sum_{n,m=1}^{N} w_{nm} \left(\sum_{i=1}^{m} c_i \, (b_i(\mathbf{x}_n) - b_i(\mathbf{x}_m)) \right)^2 \tag{8}$$

$$= (\mathbf{B}\mathbf{c} - \mathbf{y})^T \mathbf{J} \, (\mathbf{B}\mathbf{c} - \mathbf{y}) + \gamma \, \mathbf{c}^T \mathbf{B}^T \mathbf{L} \mathbf{B} \mathbf{c} \tag{9}$$

where $\mathbf{B} = (b_i(\mathbf{x}_n)) \in \mathcal{R}^{N \times m}$ can be precomputed since we fix the tree structure and parameters in all decision nodes. Again, we absorb some constants into matrices. Minimizing this over $\mathbf{c}$ yields the following linear system:

$$\mathbf{A}\mathbf{c} = \mathbf{B}^T \mathbf{J} \mathbf{y} \tag{10}$$

where $\mathbf{A} = \mathbf{B}^T \mathbf{J} \mathbf{B} + \gamma \, \mathbf{B}^T \mathbf{L} \mathbf{B}$ is the matrix of size $m \times m$ which is proportional to the number of leaves. Since oblique trees typically generate shallow trees with limited number of leaves (at most 1000 in our experiments), scalability should not be an issue. Once LapTAO is finished, *we apply the above procedure as a post-processing to the final tree*.

4 Experiments

This section shows our experimental findings. The main purpose is to demonstrate that *the proposed method dominates over other semi-supervised learning frameworks in accuracy and approaches fully supervised baseline with far less amount of labeled data*. This is true with practically no exceptions against baseline tree-based models where accuracy margin is often quite large. As for the other methods (non tree-based), we either outperform them or achieve similar error which makes LapTAO a strong competitor. To show that, we consider several regression and classification benchmarks of varying sizes and across different domains. We were able to run our algorithm on a dataset with up to 1 million instances on a regular PC which shows its scalability. Moreover, using fashion-mnist as an example, we demonstrate that the final model, a shallow oblique tree with sparse parameter vector in each node, provides insights into how it achieves a prediction allowing model interpretability.

4.1 Experimental setup

We compare our proposed approach (LapTAO) with the following baselines: 1) *oblique–all* fits an oblique tree with full supervision, this shows the theoretical maximum performance we can achieve; 2) *oblique–lbl* is the oblique trees trained on labeled portion of data $\mathcal{D}_l$ (this completely discards large portion of unlabeled data); 3) Self-training (*axis–self*, *oblique–self*) is an iterative procedure that uses the model predictions to enlarge the portion of labeled data (see section 2 for details and references), we closely follow the implementation by Yarowsky [21]; here, "axis" means traditional axis-aligned trees; 4) Laplacian SVM is a seminal work by Belkin et al. [1] which has similar problem formulation as in eq. (1) but for SVM. Additionally, the supplementary materials include comparison with semi-supervised classification trees by Levatić et al. [12] which does not change the picture much (if at all). Given the fixed cross-validation set (1% of train data), we explored as best as we could all important hyperparameters for all methods (see details in the supplementary materials). These include: controlling a tree depth (Δ), confidence threshold for self-training, σ and C values for LapSVM, etc. It worth to mention that the hyperparameter settings suggested by authors or their default values work best in most of cases.

We use TAO to train oblique trees and CART [3] to train axis-aligned trees. For all methods that use TAO, we set the total number of TAO iterations to 15. The depth Δ as well as the regularization parameter α are tuned via cross-validation. As for the settings that are specific to LapTAO, we proceed as follows. To construct the graph Laplacian, we use the Gaussian affinities with k-nearest

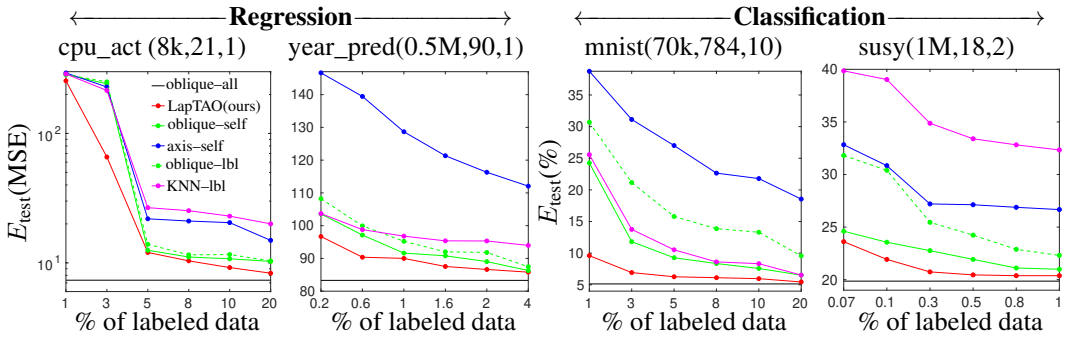

Figure 2: Results on regression (year_pred, cpu_act) and classification(mnist, susy) tasks. Numbers in brackets report the training size, number of features and output dimension (or number of classes). x–axis shows the percentage of labeled data provided to the algorithm and y–axis shows the test error. Baselines: *oblique–all* fully supervised baseline (i.e., trains an oblique tree on 100% of labeled data); *–*lbl* uses $\mathcal{D}_l$ only to train the corresponding model; *–*self* is an iterative self-training approach (see section 2 and 4.1).

neighbors and perplexity parameter $\mathcal{K}$ tuned for each dataset. The linear system in the "label–step" is solved either using direct methods (less than 20k dimensions) or Conjugate Gradient method for large scale problems. We use $\gamma = 0.1$ in all experiments. As for the main loop of the augmented Lagrangian, we iterate 20 times starting from small value for $\mu_0 = 0.001$ multiplied by 1.5 after each iteration. The remaining details as well as dataset descriptions can be found in the supplementary materials.

4.2 Main results

Fig. 2 summarizes the main results which are the trade-off plots of test error versus the percentage of labeled data on two regression and classification tasks. Intuitively, the error should go down monotonically as we increase the amount of supervision which is clearly the case in all figures. According to our findings, KNN and "axis–self" show the worst results in almost all benchmarks. The only case when KNN performed reasonably good was on MNIST which is known to work well with "template classifiers" (e.g. RBF network, kernel SVM, KNN, etc.). Even in that case it has a large error gap with respect to LapTAO. The poor performance of the "axis–self" can be explained by suboptimality of greedily grown trees [9] and suboptimality of the self-training approach which is mostly based on heuristics. Next, oblique trees trained on supervised data only ("oblique–lbl") leads to the significant drop in accuracy (magenta vs black lines). This shows that relying only on labeled data is not enough to achieve a decent performance. Incorporating an oblique tree into self-training framework brings certain benefits ("oblique–self"), notably for classification tasks (green dashed vs solid lines). Finally, *LapTAO consistently improves over all other SSL methods*, often by a considerable margin. For instance, in case of 3% in cpu_act and 1% in MNIST, the difference in the error with the second best SSL approach is several orders of magnitude. It shows acceptable results even in extreme label scarcity scenarios, e.g. when we provide $< 0.5\%$ of labeled data on year_pred and susy. Moreover, LapTAO approaches the fully supervised baseline more quickly: for MNIST, we can achieve the same $\sim 5\%$ test error as "oblique–all" using only 20% of labeled training points.

Comparison with Laplacian SVM LapSVM is a natural baseline to compare with since it uses the same problem formulation as in eq. (1) but for the support vector machines. However, we were not able to include it to the previous comparison (in fig. 2) due to implementation issues: 1) it is impractical to apply it for problems beyond 30k instances as it requires computing the inverse of the modified dense Gram matrix on the entire dataset;[1] 2) it can handle a classification task only. Therefore, we pick the subset of Fashion-MNIST (3 classes: "shirt", "bag" and "ankle boot") resulting in 18k training points. To make the comparison as fair as possible, both of the algorithms use the same graph Laplacian matrix and we enforce the same penalty on it ($\gamma = 0.1$). For LapSVM, we use the rbf kernel with $\sigma = 5$ and the hyperparameter for LIBSVM [5] is set to $C = 100$.

[1] One could approximate the inverse by Nyström (or other) method but this is beyond the scope of this paper.

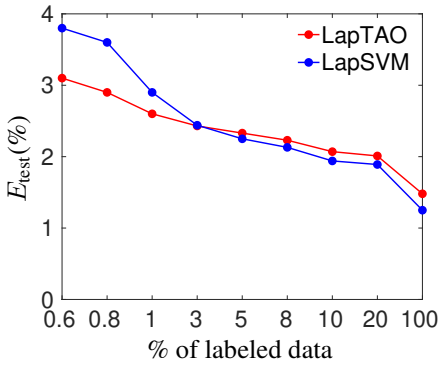

Figure 3: Comparison against LapSVM on Fashion-MNIST (3 classes: "shirt", "bag" and "ankle boot").

The results are illustrated in fig. 3. It is worth to mention that similar to the original MNIST, "template classifiers", such as kernel SVM, show quite good performance on this task which makes LapSVM a strong baseline [20]. On top of that, the problem formulation in eq. (1) is still convex for SVM and can be efficiently solved, whereas we are dealing with much harder problem for oblique trees. Therefore, it is nice to see that LapTAO performs similarly (but slightly worse) up to some point. However, it is surprising to see that the error gap between our approach and LapSVM narrows as we introduce more label scarcity and eventually we start to outperform when % of labeled data = 3%. From that point on, the table turns to the side of LapTAO and the difference becomes more and more noticeable (especially for 0.6%). One possible explanation for this behavior is the overfitting issue on small datasets which is known problem for kernel SVM, whereas sparse oblique trees are shown to be relatively robust to that [22].

Training time Overall LapTAO algorithm for the largest experiment we performed (on SUSY) took less than 24 hours and less than 2 hours for the moderate dataset size (mnist). We ran our experiments in a computer with an Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz with 32GB RAM. We did not use any GPUs.

4.3 Model interpretability

Model explainability and interpretability is a topic of renewed interest due to the widespread usage of machine learning and the risks associated with privacy, algorithmic bias, etc. In order to trust and rely on such automated systems, it is crucial to understand how they achieve a certain prediction. In contrast to "black box" models, decision trees are long considered as interpretable models due to the hierarchical structure. This allows to transform the model prediction as "if-then" rules extracted from root-to-leaf path. Specifically for oblique trees, each logical clause takes the following form: go to left child if $\mathbf{w}^T\mathbf{x} < w_0$, else go to right. This makes the interpretation a little harder since we need to look at linear combination of features at each split. However, in our case, we add ℓ_1 penalty which encourages parameter vector at each node to be sparse, i.e., only few features participate in decision making.

In this section, we argue that the oblique trees trained using LapTAO strike a good trade-off between accuracy and interpretability which is controlled via hyperparameter α. To illustrate this, we use the same subset of Fashion-MNIST dataset and train a sparse oblique tree using LapTAO (10% of training data are labeled). By decreasing the value of α we enforce more sparsity resulting into shallower and more interpretable trees (since complexity decreases). However, we sacrifice the performance since the error goes up.

Fig. 4 shows the results for $\alpha = 1$ and $\alpha = 0.1$, more results can be found in the supplementary materials. For simplicity, let us focus on the bottom tree ($\alpha = 0.1$). Clearly, each leaf contains instances of nearly the same class since the average image looks like a representative "template" from the corresponding class. As for the decision node, consider the root (node #1). All "Boot" images are sent to the right child of the root. Also, it is easy to notice that such images do not contain any pixels in the top-left quadrant of the image. Therefore, the weight vector at the root has negative (blue) values in the corresponding elements and we know that negative values are responsible for sending an instance to the left. In other words, all images that have something in the top-left quadrant are sent to the left. Therefore, boots will end up in the right child. Similarly, node #2 sends most of the images that have M-shaped stroke in the top-center part (e.g. large bags, shirts with collar) to the left child. Following the same logic, we can obtain meaningful insights for each decision node. Also note that all nodes have majority of values equal to zero (thanks to the sparsity) which makes the interpretation easier.

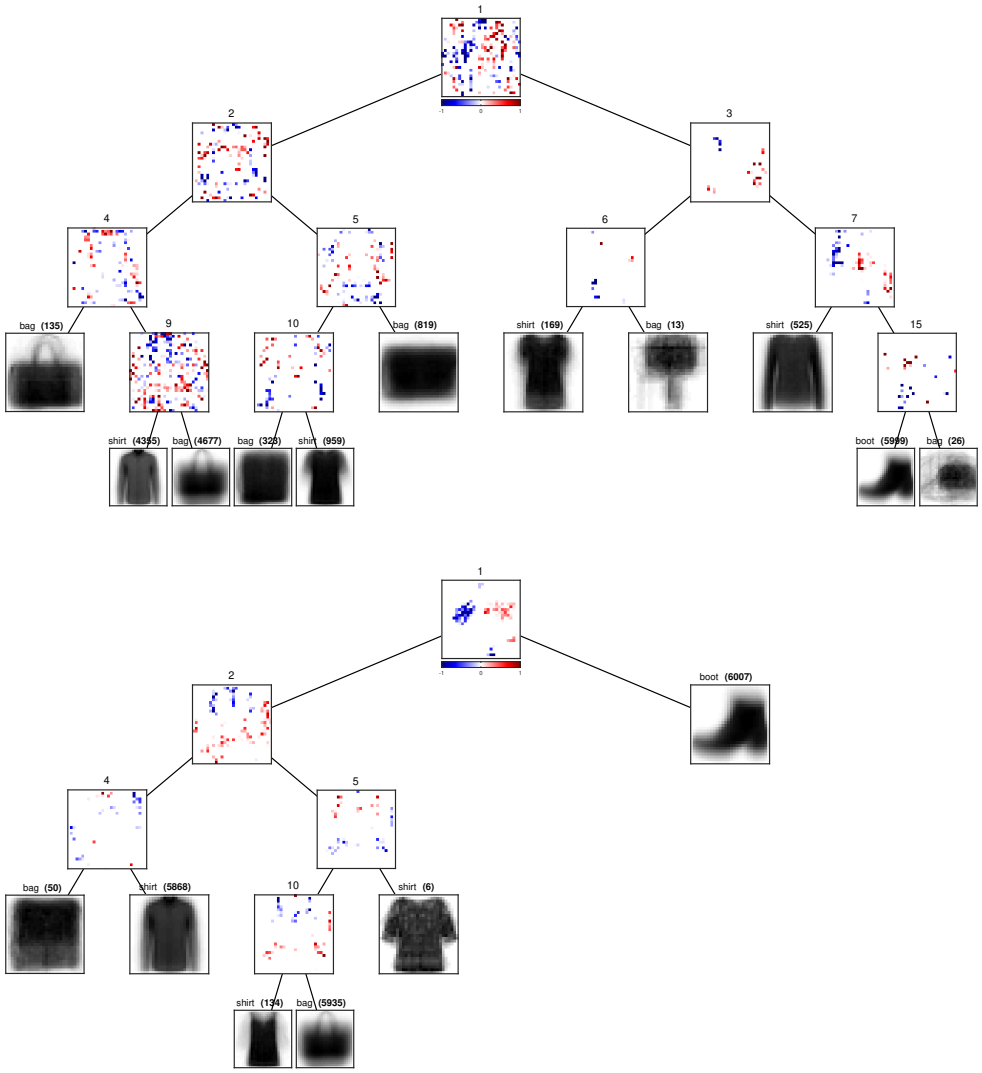

Figure 4: Some of the oblique trees obtained from LapTAO on Fashion-MNIST. Both figures use 10% of labeled data and the same graph Laplacian obtained in section 4.2, but they differ in regularization penalty (α) on the tree: (*top*) $\alpha = 1$ with $E_{\text{test}} = 2.1\%$ and (*bottom*) $\alpha = 10$ with $E_{\text{test}} = 3.9\%$. At each decision node, we illustrate the weight vector of dimension 784 reshaped into 28×28 square where each value is colored according to their sign and magnitude (positive, negative and zero values are blue, red, and white, respectively). At each leaf, we show the class label, the total number of training points in that leaf (in brackets), and the average of input images in that leaf (as a greyscale image).

5 Conclusion

We have shown how to formulate a decision tree learning problem within semi-supervised framework that involves a graph-based regularization term in an objective. The resulting problem is long considered to be intractable and we proposed an efficient iterative algorithm to minimize that objective which relies on strong optimization principles. It is based on a reformulation of the problem and decomposing it into two much simpler problems: a sparse linear system and fitting a tree. Experimental results demonstrate that the algorithm can train accurate and interpretable decision trees even in extreme label scarcity situations outperforming tree-based (and other) baselines. Our framework can generalize to other machine learning models (e.g. ensemble of trees) which will be the future direction of our work.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Machine Learning Research*, 7:2399–2434, Nov. 2006.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

[3] L. J. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, Calif., 1984.

[4] M. Á. Carreira-Perpiñán and P. Tavallali. Alternating optimization of decision trees, with application to learning sparse oblique trees. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NEURIPS)*, volume 31, pages 1211–1221. MIT Press, Cambridge, MA, 2018.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, 2(3):27, Apr. 2011.

[6] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Advances in Computer Vision and Pattern Recognition. Springer-Verlag, 2013.

[7] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *J. Machine Learning Research*, 9:1871–1874, Aug. 2008.

[8] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[9] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning—Data Mining, Inference and Prediction*. Springer Series in Statistics. Springer-Verlag, second edition, 2009.

[10] L. Hyafil and R. L. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, May 1976.

[11] C. Kemp, T. L. Griffiths, S. Stromsten, and J. B. Tenenbaum. Semi-supervised learning with trees. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 16, pages 257–264. MIT Press, Cambridge, MA, 2004.

[12] J. Levatić, M. Ceci, D. Kocev, and S. Džeroski. Semi-supervised classification trees. *J. Intelligent Information Systems*, 49:461–486, 2017.

[13] S. K. Murthy, S. Kasif, S. Salzberg, and R. Beigel. OC1: A randomized algorithm for building oblique decision trees. In *Proc. of the 11th AAAI Conference on Artificial Intelligence (AAAI 1993)*, pages 322–327, Washington, DC, July 11–15 1993.

[14] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer-Verlag, New York, second edition, 2006.

[15] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[16] J. Tanha, M. van Someren, and H. Afsarmanesh. Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*, 8:355–370, 2017.

[17] J. E. van Engelen and H. H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373–440, 2019.

[18] A. Verikas, A. Gelzinis, and M. Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349, 2011.

[19] J. Weston, F. Ratle, and R. Collobert. Deep learning via semi-supervised embedding. In A. McCallum and S. Roweis, editors, *Proc. of the 25th Int. Conf. Machine Learning (ICML'08)*, pages 1168–1175, Helsinki, Finland, July 5–9 2008.

[20] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, Sept. 15 2017.

[21] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (ACL '95)*, pages 189–196, 1995.

[22] A. Zharmagambetov, S. S. Hada, M. Gabidolla, and M. Á. Carreira-Perpiñán. Non-greedy algorithms for decision tree optimization: An experimental comparison. In *Int. J. Conf. Neural Networks (IJCNN'21)*, Virtual event, July 18–22 2021.

[23] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 16, pages 321–328. MIT Press, Cambridge, MA, 2004.

[24] X. Zhu. Semi-supervised learning literature survey. Technical Report TR–1530, Dept. of Computer Science, University of Wisconsin-Madison, July 19 2008.

[25] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU–CALD–02–107, School of Computer Science, Carnegie-Mellon University, June 2002.

[26] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In T. Fawcett and N. Mishra, editors, *Proc. of the 20th Int. Conf. Machine Learning (ICML'03)*, pages 912–919, Washington, DC, Aug. 21–24 2003.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See section 3.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is on optimization and efficiently training of machine learning models. We anticipate no impact beyond that of the models themselves.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the supplementary materials.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Section 4 and in the supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]