

Regret Bounds for Hierarchical Classification with Linear-Threshold Functions^{*}

Nicolò Cesa-Bianchi¹, Alex Conconi¹, and Claudio Gentile²

¹ Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano, Italy
`{cesa-bianchi,conconi}@dsi.unimi.it`

² Dipartimento di Informatica e Comunicazione
Università dell'Insubria, Varese, Italy
`gentile@dsi.unimi.it`

Abstract. We study the problem of classifying data in a given taxonomy when classifications associated with multiple and/or partial paths are allowed. We introduce an incremental algorithm using a linear-threshold classifier at each node of the taxonomy. These classifiers are trained and evaluated in a hierarchical top-down fashion. We then define a hierarchical and parametric data model and prove a bound on the probability that our algorithm guesses the wrong multilabel for a random instance compared to the same probability when the true model parameters are known. Our bound decreases exponentially with the number of training examples and depends in a detailed way on the interaction between the process parameters and the taxonomy structure. Preliminary experiments on real-world data provide support to our theoretical results.

1 Introduction

In this paper, we investigate the problem of classifying data based on the knowledge that the graph of dependencies between class elements is a tree forest. The trees in this forest are collectively interpreted as a taxonomy. That is, we assume that every data instance is labelled with a (possibly empty) set of class labels and, whenever an instance is labelled with a certain label i , then it is also labelled with all the labels on the path from the root of the tree where i occurs down to node i . We also allow multiple-path labellings (instances can be tagged with labels belonging to more than one path in the forest), and partial-path labellings (instances can be tagged with labels belonging to a path that does not end on a leaf).

The problem of hierarchical classification, especially of textual information, has been extensively investigated in past years (see, e.g., [5–7, 11–13, 17, 19] and references therein). Whereas the use of hierarchically trained linear-threshold classifiers is common to several of these previous approaches, to our knowledge our research is the first one to provide a rigorous performance analysis of hierarchical classification problem in the presence of multiple and partial path classifications.

^{*} The first and third author gratefully acknowledge partial support by the PASCAL Network of Excellence under EC grant no. 506778.

Following a standard approach in statistical learning theory, we assume that data are generated by a parametric and hierarchical stochastic process associated with the given taxonomy. Building on the techniques from [3], we design and analyze an algorithm for estimating the parameters of the process. Our algorithm is based on a hierarchy of regularized least-squares estimators which are incrementally updated as more data flow into the system. We prove bounds on the instantaneous regret; that is, we bound the probability that, after observing any number t of examples, our algorithm guesses the wrong multilabel on the next randomly drawn data element, while the hierarchical classifier knowing the true parameters of the process predicts the correct multilabel. Our main concern in this analysis is stressing the interaction between the taxonomy structure and the process generating the examples. This is in contrast with the standard approach in the literature about regret bounds, where a major attention is paid to studying how the regret depends on time.

To support our theoretical findings, we also briefly describe some experiments concerning a more practical variant of the algorithm we actually analyze. Though these experiments are preliminary in nature, their outcomes are fairly encouraging.

The paper is organized as follows. In Section 2 we introduce our learning model, along with the notational conventions used throughout the paper. Our hierarchical algorithm is described in Section 3 and analyzed in Section 4. In Section 5 we briefly report on the experiments. Finally, in Section 6 we summarize and mention future lines of research.

2 Learning model and notation

We assume data elements are encoded as real vectors $\mathbf{x} \in \mathbb{R}^d$ which we call *instances*. A *multilabel* for an instance \mathbf{x} is any subset of the set $\{1, \dots, c\}$ of all labels, including the empty set. We represent the multilabel of \mathbf{x} with a vector $\mathbf{v} = (v_1, \dots, v_c) \in \{-1, 1\}^c$, where i belongs to the multilabel of \mathbf{x} if and only if $v_i = 1$. A *taxonomy* G is a forest whose trees are defined over the set of labels. We use $j = \text{PAR}(i)$ to denote the unique parent of i and $\text{ANC}(i)$ to denote the set of ancestors of i . The depth of a node i (number of edges on the path from the root to i) is denoted by h_i .

A multilabel \mathbf{v} belongs to a given taxonomy if and only if it is the union of one or more paths in the forest, where each path must start from a root but need not terminate on a leaf (see Figure 1). A probability distribution f_G over the set of multilabels is associated to a taxonomy G as follows. Each node i of G is tagged with a $\{-1, 1\}$ -valued random variable V_i distributed according to a conditional probability function $\mathbb{P}(V_i \mid V_{\text{PAR}(i)}, \mathbf{X})$. To model the dependency between the labels of nodes i and $j = \text{PAR}(i)$ we assume $\mathbb{P}(V_i = 1 \mid V_j = -1, \mathbf{X} = \mathbf{x}) = 0$ for all nonroot nodes i and all instances \mathbf{x} . For example, in the taxonomy of Figure 1 we have $\mathbb{P}(V_4 = 1 \mid V_3 = -1, \mathbf{X} = \mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{R}^d$. The quantity

$$f_G(\mathbf{v} \mid \mathbf{x}) = \prod_{i=1}^c \mathbb{P}(V_i = v_i \mid V_j = v_j, j = \text{PAR}(i), \mathbf{X} = \mathbf{x})$$

thus defines a joint probability distribution on V_1, \dots, V_c conditioned on \mathbf{x} being the current instance.

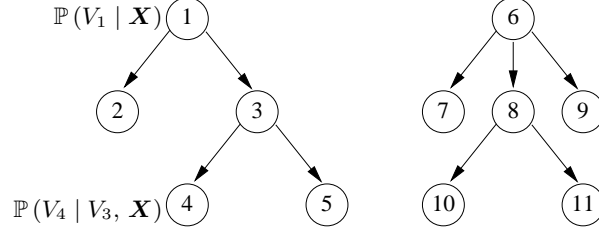


Fig. 1. A forest made up of two disjoint trees. The nodes are tagged with the name of the labels, so that in this case $c = 11$. According to our definition, the multilabel $\mathbf{v} = (1, 1, 1, -1, -1, 1, -1, 1, -1, 1, -1)$ belongs to this taxonomy (since it is the union of paths $1 \rightarrow 2$, $1 \rightarrow 3$ and $6 \rightarrow 8 \rightarrow 10$), while the multilabel $\mathbf{v} = (1, 1, -1, 1, -1, -1, -1, -1, -1, -1, -1)$ does not, since $1 \rightarrow 2 \rightarrow 4$ is not a path in the forest.

Through f_G we specify an i.i.d. process $\{(\mathbf{X}_1, \mathbf{V}_1), (\mathbf{X}_2, \mathbf{V}_2), \dots\}$, where, for $t = 1, 2, \dots$, the multilabel \mathbf{V}_t is distributed according to the joint distribution $f_G(\cdot | \mathbf{X}_t)$ and \mathbf{X}_t is distributed according to a fixed and unknown distribution D . We call each realization $(\mathbf{x}_t, \mathbf{v}_t)$ of $(\mathbf{X}_t, \mathbf{V}_t)$ an *example*.

We now introduce a parametric model for f_G . First, we assume that the support of D is the surface of the d -dimensional unit sphere (in other words, instances $\mathbf{x} \in \mathbb{R}^d$ are normalized, so that $\|\mathbf{x}\| = 1$). With each node i in the taxonomy, we associate a unit-norm weight vector $\mathbf{u}_i \in \mathbb{R}^d$. Then, we define the conditional probabilities for a nonroot node i with parent j as follows:

$$\mathbb{P}(V_i = 1 | V_j = 1, \mathbf{X} = \mathbf{x}) = (1 + \mathbf{u}_i^\top \mathbf{x})/2 . \tag{1}$$

If i is a root node, the above simplifies to

$$\mathbb{P}(V_i = 1 | \mathbf{X} = \mathbf{x}) = (1 + \mathbf{u}_i^\top \mathbf{x})/2 .$$

Note that, in this model, the labels of the children of any given node are independent random variables. This is motivated by the fact that, unlike previous investigations, we are explicitly modelling labellings involving multiple paths. A more sophisticated analysis could introduce an arbitrary negative correlation between the labels of the children nodes. We did not attempt to follow this route.

In this parametric model, we would like to perform almost as well as the hierarchical predictor that knows all vectors $\mathbf{u}_1, \dots, \mathbf{u}_c$ and labels an instance \mathbf{x} with the multilabel $\mathbf{y} = (y_1, \dots, y_c)$ computed in the following natural top-down fashion:³

$$y_i = \begin{cases} \text{SGN}(\mathbf{u}_i^\top \mathbf{x}) & \text{if } i \text{ is a root node,} \\ \text{SGN}(\mathbf{u}_i^\top \mathbf{x}) & \text{if } i \text{ is not a root and } y_j = +1 \text{ for } j = \text{PAR}(i), \\ -1 & \text{if } i \text{ is not a root and } y_j = -1 \text{ for } j = \text{PAR}(i) . \end{cases} \tag{2}$$

In other words, if a node has been labelled +1 then each child is labelled according to a linear-threshold function. On the other hand, if a node happens to be labelled -1 then *all* of its descendants are labelled -1.

³ SGN denotes the usual signum function: $\text{SGN}(x) = 1$ if $x \geq 0$; -1 , otherwise.

For our theoretical analysis, we consider the following on-line learning model. In the generic time step $t = 1, 2, \dots$ the algorithm receives an instance \mathbf{x}_t (a realization of \mathbf{X}_t) and outputs c binary predictions $\hat{y}_{1,t}, \hat{y}_{2,t}, \dots, \hat{y}_{c,t} \in \{-1, +1\}$, one for each node in the taxonomy. These predictions are viewed as guesses for the true labels $v_{1,t}, v_{2,t}, \dots, v_{c,t}$ (realizations of $V_{1,t}, V_{2,t}, \dots, V_{c,t}$, respectively) associated with \mathbf{x}_t . After each prediction, the algorithm observes the true labels and updates its estimates of the true model parameters. Such estimates will then be used in the next time step.

In a hierarchical classification framework many reasonable accuracy measures can be defined. As an attempt to be as fair as possible,⁴ we measure the accuracy of our algorithm through its global instantaneous regret on instance \mathbf{X}_t ,

$$\mathbb{P}(\exists i : \hat{y}_{i,t} \neq V_{i,t}) - \mathbb{P}(\exists i : y_{i,t} \neq V_{i,t}) ,$$

being $y_{i,t}$ the i -th label output at time t by the reference predictor (2). The above probabilities are w.r.t. the random draw of $(\mathbf{X}_1, \mathbf{V}_1), \dots, (\mathbf{X}_t, \mathbf{V}_t)$. The regret bounds we prove in Section 4 are shown to depend on the interaction between the structure of the multi-dimensional data-generating process and the structure of the taxonomy on which the process is applied.

Further notation. We denote by $\{\phi\}$ the Bernoulli random variable which is 1 if and only if predicate ϕ is true. Let ψ be another predicate. We repeatedly use simple facts such as $\{\phi \vee \psi\} = \{\phi\} + \{\psi, \neg\phi\} \leq \{\phi\} + \{\psi\}$ and $\{\phi\} = \{\phi \wedge \psi\} + \{\phi \wedge \neg\psi\} \leq \{\phi \wedge \psi\} + \{\neg\psi\}$.

3 The learning algorithm

We consider linear-threshold algorithms operating on each node of the taxonomy. The algorithm sitting on node i maintains and adjusts a weight vector $\mathbf{W}_{i,t}$ which represents an estimate at time t of the corresponding unknown vector \mathbf{u}_i .

Our hierarchical classification algorithm combines the weight vectors $\mathbf{W}_{i,t}$ associated to each node in much the same way as the hierarchical predictor (2). However, since \mathbf{u}_i parameterizes a conditional distribution where the label associated with the parent of node i is 1—recall (1), it is natural to update $\mathbf{W}_{i,t}$ only when such a conditioning event actually occurs. The pseudocode of our algorithm is given in Figure 2.

Given the i.i.d. process $\mathbf{X}_1, \mathbf{X}_2, \dots$ generating the instances, for each node i we define the derived process $\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \dots$ including all and only the instances \mathbf{X}_s of the original process that satisfy $V_{\text{PAR}(i),s} = 1$. We call this derived process the *process at node i* . Note that, for each i , the process at node i is an i.i.d. process. However, its distribution might depend on i ; that is, the process distribution at node i is generally different from the process distribution at node $j \neq i$.

⁴ It is worth mentioning that the machinery developed in this paper could also be used to analyze loss functions more sophisticated than the 0-1 loss. However, we will not pursue this more sophisticated analysis here.

Initialization: Weight vectors $\mathbf{W}_{i,1} = (0, \dots, 0)$, $i = 1, \dots, c$.
 For $t = 1, 2, \dots$ do:

1. Observe instance \mathbf{X}_t ;
2. Compute prediction values $\hat{y}_{i,t} \in \{-1, 1\}$ as follows:

$$\hat{y}_{i,t} = \begin{cases} \text{SGN}(\mathbf{W}_{i,t}^\top \mathbf{X}_t) & \text{if } i \text{ is a root node,} \\ \text{SGN}(\mathbf{W}_{i,t}^\top \mathbf{X}_t) & \text{if } i \text{ is not a root node and } \hat{y}_{j,t} = +1 \text{ for } j = \text{PAR}(i), \\ -1 & \text{if } i \text{ is not a root node and } \hat{y}_{j,t} = -1 \text{ for } j = \text{PAR}(i), \end{cases}$$

where

$$\begin{aligned} \mathbf{W}_{i,t} &= (I + S_{i,t-1} S_{i,t-1}^\top + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} S_{i,t-1} \mathbf{V}_{i,t-1} \\ \mathbf{V}_{i,t-1} &= (V_{i,i_1}, V_{i,i_2}, \dots, V_{i,i_{N(i,t-1)}})^\top \\ S_{i,t-1} &= [\mathbf{X}_{i_1} \ \mathbf{X}_{i_2} \ \dots \ \mathbf{X}_{i_{N(i,t-1)}}], \quad i = 1, \dots, c; \end{aligned}$$

3. Observe multilabel \mathbf{V}_t and perform update.

Fig. 2. The hierarchical learning algorithm.

Let $N(i, t)$ denote the number of times the *parent* of node i observes a positive label up to time t ; i.e., $N(i, t) = |\{1 \leq s \leq t : V_{\text{PAR}(i),s} = 1\}|$. The weight vector $\mathbf{W}_{i,t}$ stored at time t in node i is a (conditional) regularized least squares estimator given by

$$\mathbf{W}_{i,t} = (I + S_{i,t-1} S_{i,t-1}^\top + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} S_{i,t-1} \mathbf{V}_{i,t-1}, \quad (3)$$

where I is the $d \times d$ identity matrix, $S_{i,t-1}$ is the $d \times N(i, t-1)$ matrix whose columns are the instances $\mathbf{X}_{i_1}, \dots, \mathbf{X}_{i_{N(i,t-1)}}$ and $\mathbf{V}_{i,t-1} = (V_{i,i_1}, \dots, V_{i,i_{N(i,t-1)}})^\top$ is the $N(i, t-1)$ -dimensional vector of the corresponding labels observed by node i .

This estimator is a slight variant of regularized least squares for classification [2, 15] where we include the current instance \mathbf{x}_t in the computation of $\mathbf{W}_{i,t}$ (see, e.g., [1, 20] for analyses of similar algorithms in different contexts). Efficient incremental computations of the inverse matrix and dual variable formulations of the algorithm are extensively discussed in [2, 15].

4 Analysis

In this section we state and prove our main result, a bound on the regret of our hierarchical classification algorithm. In essence, the analysis hinges on proving that for any node i , the estimated margin $\mathbf{W}_{i,t}^\top \mathbf{X}_t$ is an asymptotically unbiased estimator of the true margin $\mathbf{u}_i^\top \mathbf{X}_t$, and then on using known large deviation arguments to obtain the stated bound. For this purpose, we bound the variance of the margin estimator at each node and prove a bound on the rate at which the bias vanishes. Both bounds will crucially depend on the convergence of the smallest empirical eigenvalue of the process at each node i , and the next result is the key to keeping this convergence under control.

Lemma 1 (Shawe-Taylor et al. [18]). *Let $\mathbf{X} = (X_1, \dots, X_d) \in \mathbb{R}^d$ be a random vector such that $\|\mathbf{X}\| = 1$ with probability 1, and let $\lambda \geq 0$ be the*

smallest eigenvalue of the correlation matrix $\{\mathbb{E}[X_i X_j]\}_{i,j=1}^d$. If $\mathbf{X}_1, \dots, \mathbf{X}_s$ are i.i.d. random vectors distributed as \mathbf{X} , S is the $d \times s$ matrix whose columns are $\mathbf{X}_1, \dots, \mathbf{X}_s$, $C = S S^\top$ is the associated empirical correlation matrix, and $\hat{\lambda}_s \geq 0$ is the smallest eigenvalue of C , then

$$\mathbb{P}\left(\frac{\hat{\lambda}_s}{s} < \lambda/2\right) \leq 2(s+1) e^{-s \lambda^2/304} \quad \text{provided } s \geq 96d/\lambda^2. \quad (4)$$

We now state our main result.

Theorem 1. Consider a taxonomy G with c nodes of depths h_1, \dots, h_c and fix an arbitrary choice of parameters $\mathbf{u}_1, \dots, \mathbf{u}_c \in \mathbb{R}^d$, such that $\|\mathbf{u}_i\| = 1$, $i = 1, \dots, c$. Assume there exist $\gamma_1, \dots, \gamma_c > 0$ such that distribution D satisfies $\mathbb{P}(\mathbf{u}_i^\top \mathbf{X}_t \geq \gamma_i) = 1$, $i = 1, \dots, c$. Then, for all

$$t > \max \left\{ \max_{i=1, \dots, c} \frac{2^{h_i+1}}{\mathbb{P}(\mathcal{A}_{i,t})} \frac{8}{\lambda_i \gamma_i}, \max_{i=1, \dots, c} \frac{2^{h_i+1}}{\mathbb{P}(\mathcal{A}_{i,t})} \frac{96d}{\lambda_i^2} \right\}$$

the regret at time t of the algorithm described in Figure 2 satisfies

$$\begin{aligned} & \mathbb{P}(\exists i : \hat{y}_{i,t} \neq V_{i,t}) - \mathbb{P}(\exists i : y_{i,t} \neq V_{i,t}) \\ & \leq \sum_{i=1}^c \mathbb{P}(\mathcal{A}_{i,t}) \left[2 e t \exp\left(-\frac{\gamma_i^2 \lambda_i (t-1) \mathbb{P}(\mathcal{A}_{i,t})}{16 \cdot 2^{h_i+1}}\right) \right. \\ & \quad \left. + e (t+1)^2 \exp\left(-\frac{\lambda_i^2 (t-1) \mathbb{P}(\mathcal{A}_{i,t})}{304 \cdot 2^{h_i+1}}\right) + \exp\left(-\frac{(t-1) \mathbb{P}(\mathcal{A}_{i,t})}{5 \cdot 2^{h_i+1}}\right) \right], \end{aligned}$$

where $\mathcal{A}_{i,t} = \{\forall j \in \text{ANC}(i) : \mathbf{u}_j^\top \mathbf{X}_t \geq 0\}$ and λ_i is the smallest eigenvalue of the process at node i .

Remark 1. Note that the dependence of $\mathbb{P}(\mathcal{A}_{i,t})$ on t is purely formal, as evinced by the definition of $\mathcal{A}_{i,t}$. Hence, the regret vanishes exponentially in t . This unnaturally fast rate is mainly caused by our assumptions on the data and, in particular, on the existence of $\gamma_1, \dots, \gamma_c$ constraining the support of D . As shown in [3], we would recover the standard $t^{-1/2}$ rate by assuming, instead, some reasonable bound on the tail of the distribution of the inverse squared margin $(\mathbf{u}_i^\top \mathbf{X}_t)^{-2}$, though this would make our analysis somewhat more complicated.

Remark 2. The values $\mathbb{P}(\mathcal{A}_{i,t})/2^{h_i}$ express the main interplay between the taxonomy structure and the process generating the examples. It is important to observe how our regret bound depends on such quantities. For instance, if we just focus on the probability values $\mathbb{P}(\mathcal{A}_{i,t})$, we see that the regret bound is essentially the sum over all nodes i in the taxonomy of terms of the form

$$\mathbb{P}(\mathcal{A}_{i,t}) \exp(-k_i \mathbb{P}(\mathcal{A}_{i,t}) t), \quad (5)$$

where the k_i 's are positive constants. Clearly, $\mathbb{P}(\mathcal{A}_{i,t})$ decreases as we descend along a path. Hence, if node i is a root then $\mathbb{P}(\mathcal{A}_{i,t})$ tends to be relatively large, whereas if i is a leaf node then $\mathbb{P}(\mathcal{A}_{i,t})$ tends to be close to zero. In both cases

(5) tends to be small: when $\mathbb{P}(\mathcal{A}_{i,t})$ is close to one it does not affect the negative exponential decrease with time; on the other hand, if $\mathbb{P}(\mathcal{A}_{i,t})$ is close to zero then (5) is small anyway. In fact, this is no surprise, since it is a direct consequence of the hierarchical nature of our prediction algorithm (Figure 2). Let us consider, for the sake of clarity, two extreme cases: 1) i is a root node; 2) i is a (very deep) leaf node.

1) A root node observes all instances. The predictor at this node is required to predict through $\text{SGN}(\mathbf{W}_{i,t}^\top \mathbf{X}_t)$ on *all* instances \mathbf{X}_t , but the estimator $\mathbf{W}_{i,t}$ gets close to \mathbf{u}_i very quickly. In this case the negative exponential convergence of the associated term (5) is fast ($\mathbb{P}(\mathcal{A}_{i,t})$ is “large”).

2) A leaf node observes a possibly small subset of the instances, but it is also required to produce only a small subset of linear-threshold predictions (the associated weight vector $\mathbf{W}_{i,t}$ might be an unreliable estimator, but is also used less often). Therefore, in this case, (5) is small just because so is $\mathbb{P}(\mathcal{A}_{i,t})$.

In summary, $\mathbb{P}(\mathcal{A}_{i,t})$ somehow measures both the rate at which the estimator in node i gets updated and the relative importance of the accuracy of this estimator when computing the overall regret.

Remark 3. The bound of Theorem 1 becomes vacuous when $\lambda_i = 0$ for some i . However, note that whenever the smallest eigenvalue of the original process (i.e., the process at the roots) is positive, then $\lambda_i > 0$ for all nodes i , up to pathological collusions between D and the \mathbf{u}_j 's. As an example of such collusions, note that the process at node i is a filtered version of the original process, as each ancestor j of i filters out \mathbf{X}_t with probability depending on the angle between \mathbf{X}_t and \mathbf{u}_j . Hence, to make the process at node i have a correlation matrix with rank strictly smaller than the one at $j = \text{PAR}(i)$, the parameter \mathbf{u}_j should be perfectly aligned with an eigenvector of the process at node j .

Remark 4. We are measuring regret against a reference predictor that is not Bayes optimal for the data model at hand. Indeed, the Bayes optimal predictor would use the maximum likelihood multilabel assignment given G and $\mathbf{u}_1, \dots, \mathbf{u}_c$ (this assignment is easily computable using a special case of the sum-product algorithm [10]). Finding a good algorithm to approximate the maximum-likelihood assignment has proven to be a difficult task.

Proof (of Theorem 1). We first observe that

$$\begin{aligned} \{\exists i : \hat{y}_{i,t} \neq V_{i,t}\} &\leq \{\exists i : y_{i,t} \neq V_{i,t}\} + \{\exists i : \hat{y}_{i,t} \neq y_{i,t}\} \\ &= \{\exists i : y_{i,t} \neq V_{i,t}\} \\ &\quad + \sum_{i=1}^c \{\hat{y}_{i,t} \neq y_{i,t}, \hat{y}_{j,t} = y_{j,t}, j = 1, \dots, i-1\}. \end{aligned} \quad (6)$$

Without loss of generality we can assume that the nodes in the taxonomy are assigned numbers such that if node i is a child of node j then $i > j$. The regret (6) can then be upper bounded as

$$\begin{aligned}
& \sum_{i=1}^c \{\hat{y}_{i,t} \neq y_{i,t}, \hat{y}_{j,t} = y_{j,t}, j = 1, \dots, i-1\} \\
& \leq \sum_{i=1}^c \{\hat{y}_{i,t} \neq y_{i,t}, \forall j \in \text{ANC}(i) : \hat{y}_{j,t} = y_{j,t}\} \\
& = \sum_{i=1}^c \{\hat{y}_{i,t} \neq y_{i,t}, \forall j \in \text{ANC}(i) : \hat{y}_{j,t} = y_{j,t} = 1\} \\
& \quad (\text{since } \hat{y}_{j,t} = y_{j,t} = -1 \text{ for some ancestor } j \text{ implies } \hat{y}_{i,t} = y_{i,t} = -1) \\
& \leq \sum_{i=1}^c \{\hat{y}_{i,t} \neq y_{i,t}, \forall j \in \text{ANC}(i) : y_{j,t} = 1\} .
\end{aligned}$$

Taking expectations we get

$$\begin{aligned}
& \mathbb{P}(\exists i : \hat{y}_{i,t} \neq V_{i,t}) - \mathbb{P}(\exists i : y_{i,t} \neq V_{i,t}) \\
& \leq \sum_{i=1}^c \mathbb{P}(\hat{y}_{i,t} \neq y_{i,t}, \forall j \in \text{ANC}(i) : y_{j,t} = 1) .
\end{aligned}$$

We now bound from above the simpler probability terms in the right-hand side. For notational brevity, in the rest of this proof we will be using $\Delta_{i,t}$ to denote the margin variable $\mathbf{u}_i^\top \mathbf{X}_t$ and $\widehat{\Delta}_{i,t}$ to denote the algorithm's margin $\mathbf{W}_{i,t}^\top \mathbf{X}_t$. As we said earlier, our argument centers on proving that for any node i , $\widehat{\Delta}_{i,t}$ is an asymptotically unbiased estimator of $\Delta_{i,t}$, and then on using known large deviation techniques to obtain the stated bound. For this purpose, we need to study both the conditional bias and the conditional variance of $\widehat{\Delta}_{i,t}$.

Recall Figure 2. We first observe that the multilabel vectors $\mathbf{V}_1, \dots, \mathbf{V}_{t-1}$ are conditionally independent given the instance vectors $\mathbf{X}_1, \dots, \mathbf{X}_{t-1}$. More precisely, we have

$$\mathbb{P}(\mathbf{V}_1, \dots, \mathbf{V}_{t-1} \mid \mathbf{X}_1, \dots, \mathbf{X}_{t-1}) = \mathbb{P}(\mathbf{V}_1 \mid \mathbf{X}_1) \times \dots \times \mathbb{P}(\mathbf{V}_{t-1} \mid \mathbf{X}_{t-1}) .$$

Also, for any given node i with parent j , the child's labels $V_{i,i_1}, \dots, V_{i,i_{N(i,t-1)}}$ are independent when conditioned on both $\mathbf{X}_1, \dots, \mathbf{X}_{t-1}$ and the parent's labels $V_{j,1}, \dots, V_{j,t-1}$. Let us denote by $\mathbb{E}_{i,t}$ the conditional expectation

$$\mathbb{E}[\cdot \mid (\mathbf{X}_1, V_{j,1}), \dots, (\mathbf{X}_{t-1}, V_{j,t-1}), \mathbf{X}_t] .$$

By definition of our parametric model (1) we have $\mathbb{E}_{i,t}[V_{i,t-1}] = S_{i,t-1}^\top \mathbf{u}_i$. Recalling the definition (3) of $\mathbf{W}_{i,t}$, this implies

$$\mathbb{E}_{i,t}[\widehat{\Delta}_{i,t}] = \mathbf{u}_i^\top S_{i,t-1} S_{i,t-1}^\top (I + S_{i,t-1} S_{i,t-1}^\top + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t .$$

In the rest of the proof, we use $\hat{\lambda}_{i,t-1}$ to denote the smallest eigenvalue of the empirical correlation matrix $S_{i,t-1} S_{i,t-1}^\top$. The conditional bias is bounded in the following lemma (proven in the appendix).

Lemma 2. *With the notation introduced so far, we have: $\Delta_{i,t} = \mathbb{E}_{i,t}[\widehat{\Delta}_{i,t}] + B_{i,t}$, where the conditional bias $B_{i,t}$ satisfies $B_{i,t} \leq 2/(1 + \hat{\lambda}_{i,t-1})$.*

Next, we consider the conditional variance of $\widehat{\Delta}_{i,t}$. Recalling Figure 2, we see that

$$\widehat{\Delta}_{i,t} = \sum_{k=1}^{N(i,t-1)} V_{i,i_k} Z_k$$

where $\mathbf{Z}^\top = (Z_1, \dots, Z_{N(i,t-1)})^\top = S_{i,t-1}^\top \left(I + S_{i,t-1} S_{i,t-1}^\top + \mathbf{X}_t \mathbf{X}_t^\top \right)^{-1} \mathbf{X}_t$. The next lemma (proven in the appendix) handles the conditional variance $\|\mathbf{Z}\|^2$.

Lemma 3. *With the notation introduced so far, we have: $\|\mathbf{Z}\|^2 \leq 1/(2 + \hat{\lambda}_{i,t-1})$.*

Armed with these two lemmas, we proceed through our large deviation argument. For the sake of brevity, denote $N(i, t-1)$ by N . Also, in order to stress the dependence of $\hat{\lambda}_{i,t-1}$, $\widehat{\Delta}_{i,t}$ and $B_{i,t}$ on $N(i, t-1)$, we denote them by $\hat{\lambda}_{i,N}$, $\widehat{\Delta}_{i,t,N}$ and $B_{i,N}$, respectively. The case when subscript N is replaced by its realization n should be intended as the random variable obtained by restricting to sample realizations such that N takes on value n . Thus, for instance, any predicate $\phi(\widehat{\Delta}_{i,t,n})$ involving $\widehat{\Delta}_{i,t,n}$ should actually be intended as a short-hand for $\phi(\widehat{\Delta}_{i,t,N}) \wedge N = n$.

Recall that $\mathcal{A}_{i,t} = \{\forall j \in \text{ANC}(i) : \mathbf{u}_j^\top \mathbf{X}_t \geq 0\} = \{\forall j \in \text{ANC}(i) : y_{j,t} = 1\}$. We have

$$\begin{aligned} & \{\hat{y}_{i,t} \neq y_{i,t}, \mathcal{A}_{i,t}\} \\ & \leq \left\{ \widehat{\Delta}_{i,t,N} - \Delta_{i,t} \leq 0, \mathcal{A}_{i,t} \right\} \\ & \leq \left\{ |\widehat{\Delta}_{i,t,N} - \Delta_{i,t}| \geq |\Delta_{i,t}|, \mathcal{A}_{i,t} \right\} \\ & \leq \left\{ |\widehat{\Delta}_{i,t,N} + B_{i,N} - \Delta_{i,t}| \geq |\Delta_{i,t}| - |B_{i,N}|, \mathcal{A}_{i,t} \right\} \\ & \leq \left\{ |\widehat{\Delta}_{i,t,N} + B_{i,N} - \Delta_{i,t}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t} \right\} + \left\{ |B_{i,N}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t} \right\}. \end{aligned} \tag{7}$$

We can bound the two terms of (7) separately. Let $M < t$ be an integer constant to be specified later. For the first term we obtain

$$\begin{aligned} & \left\{ |\widehat{\Delta}_{i,t,N} + B_{i,N} - \Delta_{i,t}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t} \right\} \\ & \leq \left\{ |\widehat{\Delta}_{i,t,N} + B_{i,N} - \Delta_{i,t}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t}, N \geq M, \hat{\lambda}_{i,N} \geq \lambda_i N/2 \right\} \\ & \quad + \left\{ \mathcal{A}_{i,t}, N \geq M, \hat{\lambda}_{i,N} < \lambda_i N/2 \right\} + \left\{ \mathcal{A}_{i,t}, N < M \right\} \\ & \leq \sum_{n=M}^{t-1} \left\{ |\widehat{\Delta}_{i,t,n} + B_{i,n} - \Delta_{i,t}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t}, \hat{\lambda}_{i,n} \geq \lambda_i n/2 \right\} \\ & \quad + \sum_{n=M}^{t-1} \left\{ \mathcal{A}_{i,t}, \hat{\lambda}_{i,n} < \lambda_i n/2 \right\} + \left\{ \mathcal{A}_{i,t}, N < M \right\}. \end{aligned}$$

For the second term, using Lemma 2 we get

$$\begin{aligned}
& \left\{ |B_{i,N}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t} \right\} \\
& \leq \left\{ \frac{2}{1 + \hat{\lambda}_{i,N}} \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t} \right\} \\
& \leq \left\{ \frac{2}{1 + \hat{\lambda}_{i,N}} \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t}, N \geq M, \hat{\lambda}_{i,N} \geq \lambda_i N/2 \right\} \\
& \quad + \left\{ \mathcal{A}_{i,t}, N \geq M, \hat{\lambda}_{i,N} < \lambda_i N/2 \right\} + \left\{ \mathcal{A}_{i,t}, N < M \right\} .
\end{aligned}$$

Now note that the choice $M \geq 8/(\lambda_i \gamma_i) \geq 8/(\lambda_i |\Delta_{i,t}|)$ makes the first term vanish. Hence, under this condition on M ,

$$\begin{aligned}
\left\{ |B_{i,N}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t} \right\} & \leq \left\{ \mathcal{A}_{i,t}, N \geq M, \hat{\lambda}_{i,N} < \lambda_i N/2 \right\} + \left\{ \mathcal{A}_{i,t}, N < M \right\} \\
& \leq \sum_{n=M}^{t-1} \left\{ \mathcal{A}_{i,t}, \hat{\lambda}_{i,n} < \lambda_i n/2 \right\} + \left\{ \mathcal{A}_{i,t}, N < M \right\} .
\end{aligned}$$

Plugging back into (7) and introducing probabilities yields

$$\begin{aligned}
& \mathbb{P}(\hat{y}_{i,t} \neq y_{i,t}, \mathcal{A}_{i,t}) \\
& \leq \sum_{n=M}^{t-1} \mathbb{P} \left(|\hat{\Delta}_{i,t,n} + B_{i,n} - \Delta_{i,t}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t}, \hat{\lambda}_{i,n} \geq \lambda_i n/2 \right) \quad (8)
\end{aligned}$$

$$+ 2 \sum_{n=M}^{t-1} \mathbb{P} \left(\mathcal{A}_{i,t}, \hat{\lambda}_{i,n} < \lambda_i n/2 \right) \quad (9)$$

$$+ 2 \mathbb{P}(\mathcal{A}_{i,t}, N < M) . \quad (10)$$

Let $j = \text{PAR}(i)$ and $\mathbb{P}_{i,t}$ denote $\mathbb{P}(\cdot \mid (\mathbf{X}_1, V_{j,1}), \dots, (\mathbf{X}_{t-1}, V_{j,t-1}), \mathbf{X}_t)$. Notice that $V_{i,i_1}, \dots, V_{i,i_{N(i,t-1)}}$ are independent w.r.t. $\mathbb{P}_{i,t}$. We bound (8) by combining Chernoff-Hoeffding inequalities [8] with Lemma 3:

$$\begin{aligned}
& \mathbb{P}_{i,t} \left(|\hat{\Delta}_{i,t,n} + B_{i,n} - \Delta_{i,t}| \geq |\Delta_{i,t}|/2, \mathcal{A}_{i,t}, \hat{\lambda}_{i,n} \geq \lambda_i n/2 \right) \\
& = \left\{ \mathcal{A}_{i,t} \right\} \times \left\{ \hat{\lambda}_{i,n} \geq \lambda_i n/2 \right\} \times \mathbb{P}_{i,t} \left(|\hat{\Delta}_{i,t,n} + B_{i,n} - \Delta_{i,t}| \geq |\Delta_{i,t}|/2 \right) \\
& \leq \left\{ \mathcal{A}_{i,t} \right\} \times \left\{ \hat{\lambda}_{i,n} \geq \lambda_i n/2 \right\} \times 2e^{-\Delta_{i,t}^2 (2 + \hat{\lambda}_{i,n})/8} \\
& \leq 2 \left\{ \mathcal{A}_{i,t} \right\} e^{-\gamma_i^2 \lambda_i n/16} .
\end{aligned}$$

Thus, integrating out the conditioning, we get that (8) is upper bounded by

$$2 \mathbb{P}(\mathcal{A}_{i,t}) \sum_{n=M}^{t-1} e^{-\gamma_i^2 \lambda_i n/16} \leq 2 \mathbb{P}(\mathcal{A}_{i,t}) t e^{-\gamma_i^2 \lambda_i M/16} .$$

Since the process at each node i is i.i.d., we can bound (9) through the concentration result contained in Lemma 1. Choosing $M \geq 96d/\lambda_i^2$, we get

$$\begin{aligned} \mathbb{P}_{i,t} \left(\mathcal{A}_{i,t}, \hat{\lambda}_{i,n} < \lambda_i n/2 \right) &= \{\mathcal{A}_{i,t}\} \mathbb{P}_{i,t} \left(\hat{\lambda}_{i,n} < \lambda_i n/2 \right) \\ &\leq 2(n+1) \{\mathcal{A}_{i,t}\} e^{-n\lambda_i^2/304}. \end{aligned}$$

Thus, integrating out the conditioning again, we get that (9) is upper bounded by

$$2 \mathbb{P}(\mathcal{A}_{i,t}) \sum_{n=M}^{t-1} (n+1) e^{-n\lambda_i^2/304} \leq \mathbb{P}(\mathcal{A}_{i,t}) (t+1)^2 e^{-M\lambda_i^2/304}.$$

Finally, we analyze (10) as follows. Recall that $N = N(i, t-1)$ counts the number of times node j , the parent of node i , has observed $V_{j,s} = 1$ for $s = 1, \dots, t-1$. Therefore $\mathbb{P}(\mathcal{A}_{i,t}, N < M) = \mathbb{P}(\mathcal{A}_{i,t}) \mathbb{P}(N < M)$, and we can focus on the latter probability. The random variable N is binomial and we can bound its parameter μ_i as follows. Let $j(1) \rightarrow j(2) \rightarrow \dots \rightarrow j(h_i) \rightarrow i$ be the unique path from a root down to node i (that is, $\text{ANC}(i) = \{j(1), \dots, j(h_i)\}$ and $j(h_i) = \text{PAR}(i)$). Fix any $\mathbf{X} \in \mathbb{R}^d$ such that $\|\mathbf{X}\| = 1$. Exploiting the way conditional probabilities are defined in our taxonomy (see Section 2), for a generic time step $s \leq t-1$ we can write

$$\begin{aligned} \mathbb{P}(V_{\text{PAR}(i),s} = 1 \mid \mathbf{X}) &= \prod_{k=1}^{h_i} \mathbb{P}(V_{j(k),s} = 1 \mid V_{j(k-1),s} = 1, \mathbf{X}) \\ &= \prod_{k=1}^{h_i} \left(\frac{1 + \mathbf{u}_{j(k)}^\top \mathbf{X}}{2} \right) \quad (\text{using (1)}) \\ &\geq \prod_{k=1}^{h_i} \left(\frac{1 + \mathbf{u}_{j(k)}^\top \mathbf{X}}{2} \right) \{\mathcal{A}_{i,t}\} \geq \left(\frac{1}{2} \right)^{h_i} \{\mathcal{A}_{i,t}\}, \end{aligned}$$

since $\mathcal{A}_{i,t}$ is equivalent to $\mathbf{u}_{j(k)}^\top \mathbf{X} \geq 0$ for $k = 1, \dots, h_i$. Integrating over \mathbf{X} we conclude that the parameter μ_i of the binomial random variable N satisfies $\mu_i = \mathbb{P}(V_{\text{PAR}(i),s} = 1) \geq \left(\frac{1}{2} \right)^{h_i} \mathbb{P}(\mathcal{A}_{i,t})$. We now set M as follows:

$$M = \lfloor (t-1)\mu_i/2 \rfloor \geq \frac{(t-1)\mathbb{P}(\mathcal{A}_{i,t})}{2^{h_i+1}} - 1.$$

This implies

$$\begin{aligned} \mathbb{P}(\mathcal{A}_{i,t}, N < M) &= \mathbb{P}(\mathcal{A}_{i,t}) \mathbb{P}(N < M) \\ &\leq \mathbb{P}(\mathcal{A}_{i,t}) e^{-(t-1)\mu_i/10} \\ &\leq \mathbb{P}(\mathcal{A}_{i,t}) \exp \left(-\frac{(t-1)\mathbb{P}(\mathcal{A}_{i,t})}{10 \cdot 2^{h_i}} \right), \end{aligned} \tag{11}$$

where we used Bernstein's inequality (see, e.g., [4, Ch. 8]) and our choice of M to prove (11).

Piecing together, overapproximating, and using in the bounds for (8) and (9) the conditions on t , along with $M \geq (t-1)\mathbb{P}(\mathcal{A}_{i,t})/2^{h_i+1} - 1$ results in

$$\begin{aligned} & \mathbb{P}(\exists i : \hat{y}_{i,t} \neq V_{i,t}) - \mathbb{P}(\exists i : y_{i,t} \neq V_{i,t}) \\ & \leq \sum_{i=1}^c \mathbb{P}(\hat{y}_{i,t} \neq y_{i,t}, \mathcal{A}_{i,t}) \\ & \leq \sum_{i=1}^c \mathbb{P}(\mathcal{A}_{i,t}) \left[2 e t \exp\left(-\frac{\gamma_i^2 \lambda_i (t-1) \mathbb{P}(\mathcal{A}_{i,t})}{16 \cdot 2^{h_i+1}}\right) \right. \\ & \quad \left. + e (t+1)^2 \exp\left(-\frac{\lambda_i^2 (t-1) \mathbb{P}(\mathcal{A}_{i,t})}{304 \cdot 2^{h_i+1}}\right) + \exp\left(-\frac{(t-1) \mathbb{P}(\mathcal{A}_{i,t})}{5 \cdot 2^{h_i+1}}\right) \right], \end{aligned}$$

thereby concluding the proof. \square

5 Preliminary experimental results

To support our theoretical results, we are testing some variants of our hierarchical classification algorithm on real-world textual data. In a preliminary series of experiments, we used the first 40,000 newswire stories from the Reuters Corpus Volume 1 (RCV1). The newswire stories in RCV1 are classified in a taxonomy of 102 nodes divided into 4 trees, where multiple-path and partial-path classifications repeatedly occur throughout the corpus. We trained our algorithm on the first 20,000 consecutive documents and tested it on the subsequent 20,000 documents (to represent documents as real vectors, we used the standard TF-IDF bag-of-words encoding — more details will be given in the full paper). To make the algorithm of Figure 2 more space-efficient, we stored in the estimator associated with each node only the examples that achieved a small margin or those that were incorrectly classified by the current estimator. In [3] this technique is shown to be quite effective in terms of the number of instances stored and not disruptive in terms of classification performance. This space-efficient version of our algorithm achieved a test error of 46.6% (recall that an instance is considered mistaken if *at least one* out of 102 labels is guessed wrong). For comparison, if we replace our estimator with the standard Perceptron algorithm [16, 14] (without touching the rest of the algorithm) the test error goes up to 65.8%, and this performance does not change significantly if we train the Perceptron algorithm at each node with all the examples independently (rather than using only the examples that are positive for the parent). For the space-efficient variant of our algorithm, we observed that training independently each node causes a moderate increase of the test error from 46.6% to 49.6%. Besides, hierarchical training is in general much faster than independent training.

6 Conclusions and ongoing research

We have introduced a new hierarchical classification algorithm working with linear-threshold functions. The algorithm has complete knowledge of the taxonomy and maintains at each node a regularized least-squares estimator of the true (unknown) margin associated to the process at that node. The predictions

at the nodes are combined in a top-down fashion. We analyzed this algorithm in the i.i.d. setting by providing a bound on the instantaneous regret, i.e., on the amount by which the probability of misclassification by the algorithm exceeds on a randomly drawn instance the probability of misclassification by the hierarchical algorithm knowing all model parameters. We also reported on preliminary experiments with a few variants of our basic algorithm.

Our analysis in Section 4 works under side assumptions about the distribution D generating the examples. We are currently investigating the extent to which it is possible to remove some of these assumptions with no further technical complications. A major theoretical open question is the comparison between our algorithm (or variants thereof) and the Bayes optimal predictor for our parametric model. Finally, we are planning to perform a more extensive experimental study on a variety of hierarchical datasets.

References

1. K.S. Azoury and M.K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.
2. N. Cesa-Bianchi, A. Conconi, and C. Gentile. A second-order Perceptron algorithm. In *Proc. 15th COLT*, pages 121–137. LNAI 2375, Springer, 2002.
3. N. Cesa-Bianchi, A. Conconi, and C. Gentile. Learning probabilistic linear-threshold classifiers via selective sampling. In *Proc. 16th COLT*, pages 373–386. LNAI 2777, Springer, 2003.
4. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, 1996.
5. S.T. Dumais and H. Chen. Hierarchical classification of web content. In *Proceedings of the 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 256–263. ACM Press, 2000.
6. M. Granitzer. *Hierarchical Text Classification using Methods from Machine Learning*. PhD thesis, Graz University of Technology, 2003.
7. T. Hofmann, L. Cai, and M. Ciaramita. Learning with taxonomies: classifying documents and words. Nips 2003: Workshop on syntax, semantics, and statistics, 2003.
8. W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
9. R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
10. F.R. Kschischang, B.J. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm *IEEE Trans. of Information Theory*, 47(2): 498–519, 2001.
11. D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proc. 14th ICML*, pages 170–178. Morgan Kaufmann Publishers, 1997.
12. A.K. McCallum, R. Rosenfeld, T.M. Mitchell, and A.Y. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *Proc. 15th ICML*, pages 359–367. Morgan Kaufmann Publishers, 1998.
13. D. Mladenic. Turning yahoo into an automatic web-page classifier. In *Proc. 13th European Conference on Artificial Intelligence*, pages 473–474, 1998.
14. A.B.J. Novikov. On convergence proofs on perceptrons. *Proc. of the Symposium on the Mathematical Theory of Automata*, vol. XII, pp. 615–622, 1962.
15. R. Rifkin, G. Yeo, and T. Poggio. Regularized least squares classification. In *Advances in Learning Theory: Methods, Model and Applications*. NATO Science

Series III: Computer and Systems Sciences, volume 190, pages 131–153. IOS Press, 2003.

16. F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386–408, 1958.
17. M.E. Ruiz and P. Srinivasan. Hierarchical text categorization using neural networks. *Information Retrieval*, 5(1):87–118, 2002.
18. J. Shawe-Taylor, C. Williams, N. Cristianini, and J.S. Kandola. On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In *Proc. 13th ALT*, pages 23–40. LNCS 2533, Springer, 2002.
19. A. Sun and E.-P. Lim. Hierarchical text classification and evaluation. In *Proc. 2001 International Conference on Data Mining*, pages 521–528. IEEE Press, 2001.
20. V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

Appendix

This appendix contains the proofs of Lemma 2 and Lemma 3 mentioned in the main text. Recall that, given a positive definite matrix A , the spectral norm of A , denoted by $\|A\|$, equals the largest eigenvalue of A . As a simple consequence, $\|A^{-1}\|$ is the reciprocal of the smallest eigenvalue of A .

Proof of Lemma 2

Setting $A = I + S_{i,t-1}S_{i,t-1}^\top$ we get

$$\begin{aligned}\Delta_{i,t} &= \mathbb{E}_{i,t}[\widehat{\Delta}_{i,t}] + \mathbf{u}_i^\top (I + \mathbf{X}_t \mathbf{X}_t^\top) (A + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t \\ &= \mathbb{E}_{i,t}[\widehat{\Delta}_{i,t}] + \mathbf{u}_i^\top (A + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t + \Delta_{i,t} \mathbf{X}_t^\top (A + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t.\end{aligned}\quad (12)$$

Using the Sherman-Morrison formula (e.g., [9, Ch. 1]) and the symmetry of A , we can rewrite the second term of (12) as

$$\begin{aligned}\mathbf{u}_i^\top (A + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t &= \mathbf{u}_i^\top \left(A^{-1} - \frac{A^{-1} \mathbf{X}_t \mathbf{X}_t^\top A^{-1}}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} \right) \mathbf{X}_t \\ &= \mathbf{u}_i^\top A^{-1} \mathbf{X}_t - \frac{\mathbf{u}_i^\top A^{-1} \mathbf{X}_t \mathbf{X}_t^\top A^{-1} \mathbf{X}_t}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} = \frac{\mathbf{u}_i^\top A^{-1} \mathbf{X}_t}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t}\end{aligned}$$

and the third term of (12) as

$$\Delta_{i,t} \mathbf{X}_t^\top (A + \mathbf{X}_t \mathbf{X}_t^\top)^{-1} \mathbf{X}_t = \Delta_{i,t} \frac{\mathbf{X}_t^\top A^{-1} \mathbf{X}_t}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t}.$$

Plugging back into (12) yields $\Delta_{i,t} = \mathbb{E}_{i,t}[\widehat{\Delta}_{i,t}] + B_{i,t}$ where the conditional bias $B_{i,t}$ satisfies

$$\begin{aligned}B_{i,t} &= \frac{\mathbf{u}_i^\top A^{-1} \mathbf{X}_t}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} + \Delta_{i,t} \frac{\mathbf{X}_t^\top A^{-1} \mathbf{X}_t}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} \\ &\leq \frac{\|\mathbf{u}_i\| \|A^{-1}\| \|\mathbf{X}_t\|}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} + \frac{|\Delta_{i,t}| \|\mathbf{X}_t\|^2 \|A^{-1}\|}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} \\ &\leq \frac{\|A^{-1}\|}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} + \frac{\|A^{-1}\|}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} \leq 2\|A^{-1}\|.\end{aligned}$$

Here the second inequality holds because $\|\mathbf{u}_i\| = \|\mathbf{X}_t\| = 1$ and $|\Delta_{i,t}| \leq \|\mathbf{u}_i\| \|\mathbf{X}_t\| = 1$, and the third inequality holds because $\mathbf{X}_t^\top A^{-1} \mathbf{X}_t \geq 0$ by the positive definiteness of A^{-1} . Recalling that $\|A^{-1}\| = 1/(1 + \hat{\lambda}_{i,t-1})$, where $1 + \hat{\lambda}_{i,t-1}$ is the smallest eigenvalue of A , concludes the proof. \square

Proof of Lemma 3

Setting for brevity $H = S_{i,t-1}^\top A^{-1} \mathbf{X}_t$ and $r = \mathbf{X}_t^\top A^{-1} \mathbf{X}_t$ we can write

$$\begin{aligned}
\|\mathbf{Z}\|^2 &= \mathbf{X}_t^\top \left(A + \mathbf{X}_t \mathbf{X}_t^\top \right)^{-1} S_{i,t-1} S_{i,t-1}^\top \left(A + \mathbf{X}_t \mathbf{X}_t^\top \right)^{-1} \mathbf{X}_t \\
&= \mathbf{X}_t^\top \left(A^{-1} - \frac{A^{-1} \mathbf{X}_t \mathbf{X}_t^\top A^{-1}}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} \right) S_{i,t-1} S_{i,t-1}^\top \left(A^{-1} - \frac{A^{-1} \mathbf{X}_t \mathbf{X}_t^\top A^{-1}}{1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t} \right) \mathbf{X}_t \\
&\quad (\text{by the Sherman-Morrison formula}) \\
&= H^\top H - \frac{r}{1+r} H^\top H - \frac{r}{1+r} H^\top H + \frac{r^2}{(1+r)^2} H^\top H \\
&= \frac{H^\top H}{(1+r)^2} = \frac{\mathbf{X}_t^\top A^{-1} S_{i,t-1} S_{i,t-1}^\top A^{-1} \mathbf{X}_t}{\left(1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t\right)^2} \\
&\leq \frac{\|A^{-1/2} \mathbf{X}_t\| \|A^{-1/2} S_{i,t-1} S_{i,t-1}^\top A^{-1/2}\| \|\mathbf{X}_t^\top A^{-1/2}\|}{\left(1 + \mathbf{X}_t^\top A^{-1} \mathbf{X}_t\right)^2} \\
&= \frac{r}{(1+r)^2} \|A^{-1/2} S_{i,t-1} S_{i,t-1}^\top A^{-1/2}\|. \tag{13}
\end{aligned}$$

We continue by bounding the two factors in (13). Observe that

$$r = \mathbf{X}_t^\top A^{-1} \mathbf{X}_t \leq \|A^{-1}\| = \frac{1}{1 + \hat{\lambda}_{i,t-1}} \leq 1$$

and that the function $f(x) = x/(1+x)^2$ is monotonically increasing when $x \in [0, 1]$. Hence

$$\frac{r}{(1+r)^2} = f(r) \leq f\left(\frac{1}{1 + \hat{\lambda}_{i,t-1}}\right) = \frac{1 + \hat{\lambda}_{i,t-1}}{(2 + \hat{\lambda}_{i,t-1})^2} \leq \frac{1}{2 + \hat{\lambda}_{i,t-1}}.$$

As far as the second factor is concerned, we just note that the two matrices $A^{-1/2}$ and $S_{i,t-1} S_{i,t-1}^\top$ have the same eigenvectors. Therefore

$$\left\| A^{-1/2} S_{i,t-1} S_{i,t-1}^\top A^{-1/2} \right\| = \frac{\hat{\lambda}}{1 + \hat{\lambda}} \leq 1,$$

where $\hat{\lambda}$ is *some* eigenvalue of $S_{i,t-1} S_{i,t-1}^\top$. Substituting into (13) yields

$$\|\mathbf{Z}\|^2 \leq \frac{1}{2 + \hat{\lambda}_{i,t-1}},$$

as desired. \square