Absolute convergence and error thresholds in non-active adaptive sampling

Manuel Vilares Ferro [a,*], Víctor M. Darriba Bilbao [a], Jesús Vilares Ferro [b]

[a] *Department of Computer Science, University of Vigo, Campus As Lagoas s/n, 32004 – Ourense, Spain*
[b] *Department of Computer Science, University of A Coruña, Campus de Elviña, 15071 – A Coruña, Spain*

ARTICLE INFO

ABSTRACT

Non-active adaptive sampling is a way of building machine learning models from a training data base which are supposed to dynamically and automatically derive guaranteed sample size. In this context and regardless of the strategy used in both scheduling and generating of weak predictors, a proposal for calculating absolute convergence and error thresholds is described. We not only make it possible to establish when the quality of the model no longer increases, but also supplies a proximity condition to estimate in absolute terms how close it is to achieving such a goal, thus supporting decision making for fine-tuning learning parameters in model selection. The technique proves its correctness and completeness with respect to our working hypotheses, in addition to strengthening the robustness of the sampling scheme. Tests meet our expectations and illustrate the proposal in the domain of natural language processing, taking the generation of part-of-speech taggers as case study.

© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

A recurrent issue in *machine learning* (ML) relates the determination of optimal sampling data sets, the aim being to reduce both training costs and time without making the modelling process less reliable. In this sense, the operating principle for adaptive sampling is simple and involves beginning with an initial number of examples and then iteratively learning the model, evaluating it and acquiring additional observations if necessary. Accordingly, there are two questions to be considered: it is necessary to determine the training data to be acquired at each cycle, and also to define a halting condition to terminate the loop once a certain degree of performance has been achieved by the learner. Both tasks confer the character of research issues to the formalization of *scheduling* and *stopping criteria* [20], respectively. The former has been researched for decades in terms of fixed [20,30] or adaptive [30] sequencing, and it is not our objective. As regards the halting criteria, they are independent of the scheduling and mostly start from the hypothesis that learning curves are well-behaved, including an initial steeply sloping portion, a more gently sloping middle one and a final balanced zone [28]. Accordingly, the purpose is to identify the moment in which such a curve reaches a plateau, namely when adding more data instances does not improve the accuracy, although this often does not strictly verify. Instead, extra learning efforts almost always result in modest increases. This justifies the interest in having a *proximity condition*, understood as a measure of the degree of convergence attained from a given iteration, rather than a stopping one. In short, this will make it possible to select the level of reliability in predicting a learner's performance, both in terms of accuracy and computational costs. We will thus

have a powerful and flexible decision support tool in the field of model selection, capable of adapting to the user's needs in terms of the evaluation quality of both the learning strategy and its parameterization.

A major challenge is then to avoid the overvaluation of learning perturbations, in such a way that the training does not stop prematurely due to temporary increases in accuracy. Namely, we are interested in proving the *correctness* of a proximity criterion with respect to the working hypotheses, but also in improving its capacity to mitigate the impact of such fluctuations without compromising it, i.e. its *robustness*. Given that we are looking for a practical formula, it is finally necessary to ensure its applicability, which relies on proving the *completeness* of the approach. These properties focus the attention of this work, the structure of which we briefly describe. Firstly, Section 2 examines the methodologies serving as an inspiration to solve the question posed, as well as our contributions. Next, Section 3 reviews the mathematical basis supporting the proposal, whose model we introduce in Section 4. In Section 5, we describe the testing frame for the experiments illustrated in Section 6. Finally, Section 7 presents the final conclusions.

2. The state of the art

Below is a brief review on how correctness, robustness and completeness have been addressed over time in the definition of halting conditions in adaptive sampling, thus allowing to contextualize our contribution in that respect.

2.1. Working on correctness, robustness and completeness

Regarding correctness, most adaptive samplers assume a set of hypotheses guaranteeing concurrence, such as access to independent and identically distributed observations [36,38]. The learning curve is then monotonic and, since it is bounded, training converges on a supremum. At this point, the conditions for halting are addressed from two viewpoints, depending on whether predictive accuracy is the only factor to take into account [14] or just another one in an optimization scenario stated in *decision theory* [19]. In this latter context, performance is understood as the search for a proper cost/benefit trade-off and authors resort to statistically based strategies by applying the principle of *maximum expected utility* [28] (MEU). This implies taking all effectiveness considerations into account, which depends on the degree of control exercised by the user on the learning process. In its absence, namely using non-active techniques as we do, the final cost is the sum of data acquisition, error and model induction charges [44]. Nonetheless, at best, heuristics are used to calculate the first two and there is thus no way of guaranteeing the location of a global optimum [22], which often results in assuming fixed budgets [21]. Alternatively, procedures exclusively based on accuracy estimates try to identify the plateau of the learning curve in terms of functional convergence. Among the most popular ones are *local detection* and *learning curve estimation* [20], or *linear regression with local sampling* [30], all of them based on heuristics. Again, we cannot talk here about proximity criteria, only of stopping conditions. More recently, this issue has been corrected [42], although the proposal is still far from our objective because the proximity is expressed in terms of the net contribution of each iteration to the learning process, which provides not absolute but relative estimates.

Turning to robustness, one common idea is to generate different versions (*weak predictors*) of the partial learning curves by changing the training data distribution repeatedly, and integrating the hypotheses thus obtained. That way, *bagging*[1] procedures [6] build the predictors in parallel to combine them by voting (*classification*) [24] or averaging (*regression*) [23]. On the contrary, *boosting* algorithms [33] do it sequentially, which allows the adapting of such a distribution from the results observed in previous predictors. This gives rise to *arcing*[2] strategies [13], where increasing weight is placed on the more frequently misclassified observations. Since these are the troublesome points, focusing on them may do better than the neutral bagging approach [2], justifying [15] its popularity. Another well-known method is the *k-fold cross validation* [9], where the sample is randomly partitioned into k equal sized subsamples. For each fold, a model is trained on the other $k-1$ ones and tested on it, which gives an advantage to working with small data sets. The performance reported is the average of the values computed. Whatever the format, such as online proposals, all these build on the observations available, a major constraint for making estimations beyond the last one. One simple way to alleviate this problem is by using *anchors* [42], i.e. extra examples placed at the point of infinity to generate the weak predictor in each cycle. As any one of such curves is the result of a fitting action, the sum total of its *residuals*, namely the differences between the observed values and the fitted ones, is null. This gives the anchor the chance to neutralize irregularities by choosing an appropriate value.

Finally, completeness of the halting conditions has received no attention before to the best of our knowledge, probably because so far no additional assumptions on the sampling premises were necessary to provide a practical solution.

2.2. Our contribution

It revolves around foundations, reliability and applicability to provide correctness, robustness and completeness in a context for which the ease of use is a priority. The former is established from a set of working hypotheses widely recognized in ML and a previous outcome, whose interest was only formal to date, on learning convergence in adaptive sampling [42].

---

[1] For *bootstrap aggregating.*

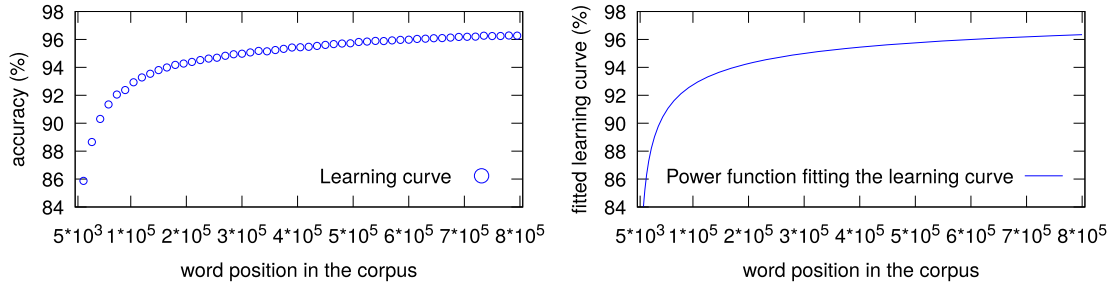[2] For *adaptive resampling and combining.*

Fig. 1. Learning curve for fnTBL on FROWN corpus, and an accuracy pattern fitting it.

To that end, the adaptation to the premises of the theoretical result must be guaranteed. The solution is based on the concept of *anchoring*, also proposed by those authors but only as a robustness mechanism, which requires the development of a specific family of such techniques.

We choose the domain of *natural language processing* (NLP) as case study, and more precisely the modelling of *part-of-speech* (POS) *taggers*, the classifiers that mark a word in a text (corpus) as corresponding to a particular POS,[3] based on its definition and context. The reasons are the significant resource and time costs of generating training data, the complexity of the relations to be learned and the fact that POS tagging is prior to any other NLP task, so errors at this stage can lower its performance [37]. That highlights the scale of the challenge, but also justifies its interest and popularity as experimentation field for new ML facilities, particularly around sampling technology [3,25,32,35,43], which is the case here.

3. The formal framework

We introduce the concepts underlying the proposal, most of them taken from [42], denoting the real numbers by $\mathbb{R}$ and the natural ones by $\mathbb{N}$, assuming that $0 \notin \mathbb{N}$. A prior question to clarify, because the generation of ML-based POS taggers serves as an illustration guide, is the accuracy notion usually accepted in that kind of model. We define it as the number of correctly tagged tokens divided by the total ones, expressed as a percentage [41] and calculated following some generally admitted usages: all tokens in the testing data set are counted, including punctuation marks, and it is supposed that only one tag *per* token is provided.

3.1. The working hypotheses

We start with a sequence of observations calculated from cases incrementally taken from a training data base, meeting some conditions to ensure a predictable progression of the estimates over a virtually infinite interval. So, they are assumed to be independently and identically distributed [12,36,38]. We then accept that a learning curve is a positive definite and strictly increasing function on $\mathbb{N}$, where numbers are the position of the case in the training data base, and upper bounded by 100. This results [1] in a concave graph with horizontal asymptote. Such hypotheses make up an idealized working frame to support correctness, while real learners may deviate from it, thus justifying the study of robustness. These deviations impact both the concavity and increase of those curves, as shown in the left-most diagram of Fig. 1 for the *fast transformation-based learning* (fnTBL) tagger [29] on the *Freiburg-Brown* (FROWN) corpus of American English [26].

3.2. The notational support

Having identified the context of the problem, it is necessary to formalize the data structures we are going to work with, such as the collection of instances whose convergence is intended to be measured.

Definition 1. (Learning scheme) *Let $\mathcal{D}$ be a training data base, $\mathcal{K} \subsetneq \mathcal{D}$ a set of initial items from $\mathcal{D}$, and $\sigma : \mathbb{N} \to \mathbb{N}$ a function. We define a* learning scheme *for $\mathcal{D}$ with* kernel $\mathcal{K}$ *and* step σ, *as a triple $\mathcal{D}_\sigma^\mathcal{K} = [\mathcal{K}, \sigma, \{\mathcal{D}_i\}_{i\in\mathbb{N}}]$, such that $\{\mathcal{D}_i\}_{i\in\mathbb{N}}$ is a cover of $\mathcal{D}$ verifying:*

$$\mathcal{D}_1 := \mathcal{K} \text{ and } \mathcal{D}_i := \mathcal{D}_{i-1} \cup \mathcal{I}_i, \ \mathcal{I}_i \subset \mathcal{D} \setminus \mathcal{D}_{i-1}, \ \|\mathcal{I}_i\| = \sigma(i), \ \forall i \geq 2 \tag{1}$$

with $\|\mathcal{I}_i\|$ the cardinality of $\mathcal{I}_i$. We refer to $\mathcal{D}_i$ as the individual of level *i for $\mathcal{D}_\sigma^\mathcal{K}$.*

---

[3] A POS is a class of words which have similar grammatical properties. Words that are assigned to the same POS generally display similar behaviour in terms of syntax, i.e. they play analogous roles within the grammatical structure of sentences. The same applies in terms of morphology, in that they undergo inflection for similar properties. Common POS labels are lexical categories (noun, verb, adjective, adverb, pronoun, preposition, conjunction, interjection, numeral, article, determiner, ...), the number or the gender.
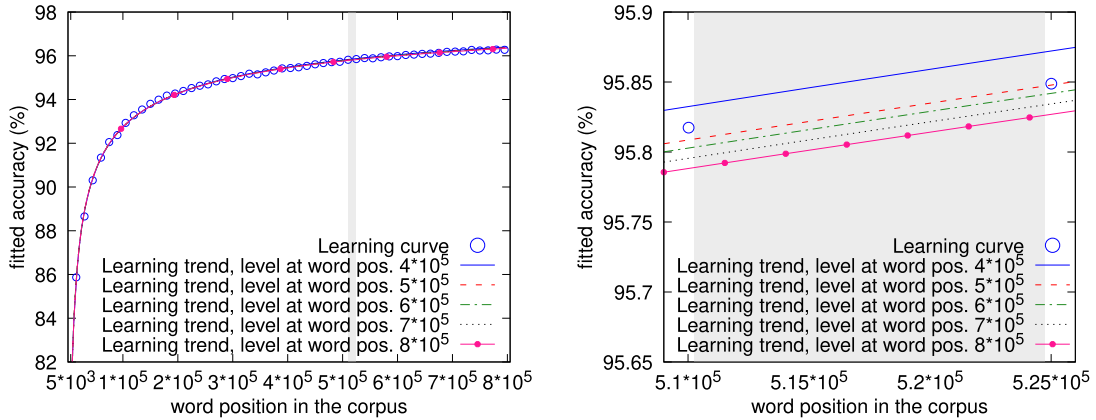
Fig. 2. Learning trace for fnTBL on FROWN, with details in zoom.

A learning scheme relates a level i with the position $\|\mathcal{D}_i\|$ in the training data base, determining the sequence of observations $\{[x_i, \mathcal{A}_\infty[\mathcal{D}](x_i)], \; x_i := \|\mathcal{D}_i\|\}_{i \in \mathbb{N}}$, where $\mathcal{A}_\infty[\mathcal{D}](x_i)$ is the accuracy achieved on such an instance by the learner. Thus, a level determines an iteration in the adaptive sampling whose learning curve is $\mathcal{A}_\infty[\mathcal{D}]$, whilst $\mathcal{K}$ delimits a portion of $\mathcal{D}$ we believe to be enough to initiate consistent evaluations of the training. For its part, σ identifies the sampling scheduling. As we want stable estimates, partial learning curves are extrapolated according to a functional pattern that verifies the working hypotheses, but are also infinitely differentiable. This supplies graphs without disruptions due to instantaneous jumps while ensuring their regularity.

Definition 2. (Accuracy pattern) *Let $C_{(0,\infty)}^\infty$ be the C-infinity functions in $\mathbb{R}^+$, we say that $\pi : \mathbb{R}^{+^n} \to C_{(0,\infty)}^\infty$ is an accuracy pattern iff $\pi(a_1, \ldots, a_n)$ is positive definite, upper bounded, concave and strictly increasing.*

An example is the *power law family* of curves $\pi(a, b, c)(x) := -a * x^{-b} + c$, hereafter used as the running one. Its upper bound is the horizontal asymptote value $\lim_{x \to \infty} \pi(a, b, c)(x) = c$, and

$$\pi(a, b, c)'(x) = a * b * x^{-(b+1)} > 0 \qquad \pi(a, b, c)''(x) = -a * b * (b+1) * x^{-(b+2)} < 0 \tag{2}$$

which guarantees increase and concavity in $\mathbb{R}^+$, respectively. This is illustrated in the right-most diagram of Fig. 1, whose goal is to fit the learning curve represented on the left-hand side. Here, the values $a = 542.5451$, $b = 0.3838$ and $c = 99.2876$ are provided by the *trust region method* [4], a regression technique minimizing the summed square of *residuals*, namely the differences between the observed values and the fitted ones. Furthermore, as the aim is to determine the degree of convergence attained by the learning process, we need to evaluate the progression of accuracy through the sequence of weak predictors being computed.

Definition 3. (Learning trend) *Let $\mathcal{D}_\sigma^\mathcal{K}$ be a learning scheme, π an accuracy pattern and $\ell \in \mathbb{N}$, $\ell \geq 3$ a position in the training data base $\mathcal{D}$. We define the* learning trend *of level ℓ for $\mathcal{D}_\sigma^\mathcal{K}$ using π, as a curve $\mathcal{A}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}] \in \pi$, fitting the observations $\{[x_i, \mathcal{A}_\infty[\mathcal{D}](x_i)], \; x_i := \|\mathcal{D}_i\|\}_{i=1}^\ell$. A sequence of learning trends $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}] := \{\mathcal{A}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}]\}_{\ell \in \mathbb{N}}$ is called a* learning trace. *We refer to $\{\alpha_\ell\}_{\ell \in \mathbb{N}}$ as the* asymptotic backbone *of $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$, where $y = \alpha_\ell := \lim_{x \to \infty} \mathcal{A}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}](x)$ is the asymptote of $\mathcal{A}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}]$.*

A learning trend $\mathcal{A}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ requires a level $\ell \geq 3$, because we need at least three observations to generate a curve. Its value $\mathcal{A}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}](x_i)$ represents the prediction for accuracy on a case x_i, using a model generated from the first ℓ cycles of the learner. Accordingly, the asymptotic term α_ℓ is nothing other than the estimate for the highest accuracy attainable. This way, a learning trace gives a comprehensive picture of the increase in accuracy due to new observations, as well as future expectations. Continuing with the tagger fnTBL and the corpus FROWN, Fig. 2 illustrates a portion of the learning trace with kernel and uniform step function $5 * 10^3$, including a zoom view.

4. The abstract model

Learning traces lay the foundations for estimating absolute convergence and error thresholds in adaptive sampling [42], thus giving coverage to the correctness we are looking for, while only a relative solution is described in practice. In short, assessing is done from the gain of accuracy between consecutive iterations, which is not enough for our purposes. To overcome this limit, we turn to the concept of *anchoring*, originally introduced to improve robustness, but which is now also useful to ensure completeness. The problem formulates in terms of the uniform approximation of a learning curve

$\mathcal{A}_{\infty}[\mathcal{D}]$ by means of the limit function $\mathcal{A}_{\infty}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ for a learning trace $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}] := \{\mathcal{A}_{i}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]\}_{i\in\mathbb{N}}$ incrementally built from sampling. We start with a brief reminder of the key results on robustness and correctness. The reader can focus on the less formal aspects, to later address in detail completeness as main contribution.

4.1. Robustness

Real learning conditions may diverge slightly from the ideal ones in the working hypotheses on which correctness is stated. In this sense, robustness is studied in the context of a more flexible set of *testing hypotheses*. These captures the notion of irregular observation by assuming that learning curves are positive definite and upper bounded by 100, but only quasi-strictly increasing and concave. We then differentiate the alterations according to their position in relation to the *working level* (wlevel), i.e. the cycle from which they would have a small enough impact to work on their softening. As this depends on unpredictable factors such as the magnitude, distribution and the very existence of these disorders, a heuristic is necessary to identify it. Considering that the model stabilizes as the training advances and that the monotony of the asymptotic backbone is at the basis of the correctness for any halting condition, a way of doing it is to categorize the variations induced in the latter. This allows to estimate wlevel as the level providing the first fluctuation below a given ceiling and, once passed, the *prediction level* (plevel) marking the beginning for learning trends which could feasibly predict the learning curve, namely not exceeding its maximum (100).

Definition 4. (Working and prediction levels) *Let $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ be a learning trace with asymptotic backbone $\{\alpha_i\}_{i\in\mathbb{N}}$, $\nu \in (0,1)$, $\varsigma \in \mathbb{N}$ and $\lambda \in \mathbb{N} \cup \{0\}$. The* working level (wlevel) *for $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ with verticality threshold ν, slowdown ς and look-ahead λ, is the smallest $\omega(\nu, \varsigma, \lambda) \in \mathbb{N}$ verifying*

$$\frac{\sqrt[\varsigma]{\nu}}{1-\nu} \geq \frac{|\alpha_{i+1} - \alpha_i|}{x_{i+1} - x_i}, \ x_i := \|\mathcal{D}_i\|, \ \forall i \text{ such that } \omega(\nu, \varsigma, \lambda) \leq i \leq \omega(\nu, \varsigma, \lambda) + \lambda \tag{3}$$

while the smallest $\wp(\nu, \varsigma, \lambda) \geq \omega(\nu, \varsigma, \lambda)$ with $\alpha_{\wp(\nu, \varsigma, \lambda)} \leq 100$ is the prediction level (plevel). *Unless they are necessary for understanding, we shall omit the parameters, referring to* wlevel *by ω (resp.* plevel *by $\wp$).*

The wlevel is the first level for which the normalized absolute value of the slope of the line joining consecutive points on the asymptotic backbone is less than the verticality threshold ν, which is corrected by a factor $1/\varsigma$ in order to slow down the normalization pace for ν. In effect, since the absolute value for a slope is defined in the interval $[0, \infty)$, the normalizing function to be applied can be expressed as follows:

$$\begin{array}{rcl} \text{N}: [0, \infty) & \longrightarrow & [0, 1) \\ x & \rightsquigarrow & \frac{x}{x-1} \end{array} \tag{4}$$

That way, given two consecutive points (x_i, α_i) and (x_{i+1}, α_{i+1}) in the asymptotic backbone, the absolute slope to be considered and the original condition we are looking for are then, respectively:

$$\frac{|\alpha_{i+1} - \alpha_i|}{x_{i+1} - x_i} \qquad \text{and} \qquad \text{N}\left(\frac{|\alpha_{i+1} - \alpha_i|}{x_{i+1} - x_i}\right) < \nu \tag{5}$$

The latter condition can be easily transformed into the following equivalent one:

$$\frac{\nu}{1-\nu} \geq \frac{|\alpha_{i+1} - \alpha_i|}{x_{i+1} - x_i} \tag{6}$$

from which, including the slowdown factor $1/\varsigma$ and the condition on the look-ahead λ, we derive Eq. (3). The use of normalized slopes corrected by the slowdown parameter ς allows recursion to infinitely large values and to extremely small decimal fractions to be avoided, thus facilitating the setting of the threshold ν.

Intuitively, since slope values decrease together with the deviations in the monotony studied, we can use them to categorize the latter, taking the look-ahead λ as verification window. We then place plevel on the first cycle with a learning trend below 100, which would therefore be the first level with real predictive capacity, since the previous ones would exceed this maximum accuracy value for any model generated. In our example, Fig. 3 shows the scale of such deviations before and after wlevel, for $\nu = 2 * 10^{-5}$, $\varsigma = 1$ and $\lambda = 5$. Now the way is clear to introduce *anchoring* as a mechanism for robustness in sampling.

Definition 5. (Anchoring learning trace) *Let $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ be a learning trace with* wlevel ω, *and $\{\hat{\mathcal{A}}_{\ell}(\infty)\}_{\ell > \omega} \subset \mathbb{R}^+$. A learning trend of level $\ell > \omega$ with anchor $\hat{\mathcal{A}}_{\ell}(\infty)$ for $\mathcal{A}_{\infty}[\mathcal{D}]$ using the accuracy pattern π, is a curve $\hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}] \in \pi$ fitting the observations $\{[x_i, \mathcal{A}_{\infty}[\mathcal{D}](x_i)], x_i := \|\mathcal{D}_i\|\}_{i=1}^{\ell} \cup [\infty, \hat{\mathcal{A}}_{\ell}(\infty)]$. We denote by $\hat{\rho}_{\ell}(i) := [\mathcal{A}_{\infty}[\mathcal{D}] - \hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]](x_i)$ the residual of $\hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ at the level i, by $\hat{\rho}_{\ell}(\infty) := \hat{\mathcal{A}}_{\ell}(\infty) - \hat{\alpha}_{\ell}$ its residual at the point of infinity and by $y = \hat{\alpha}_{\ell}$ its asymptote. When $\{\hat{\alpha}_{\ell}\}_{\ell > \omega}$ is positive definite and converges monotonically to the asymptotic value α_{∞} of $\mathcal{A}_{\infty}[\mathcal{D}]$, we say that $\hat{\mathcal{A}}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}] := \{\hat{\mathcal{A}}_{\ell}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]\}_{\ell > \omega}$ is an anchoring learning trace of reference $[\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \omega]$.*
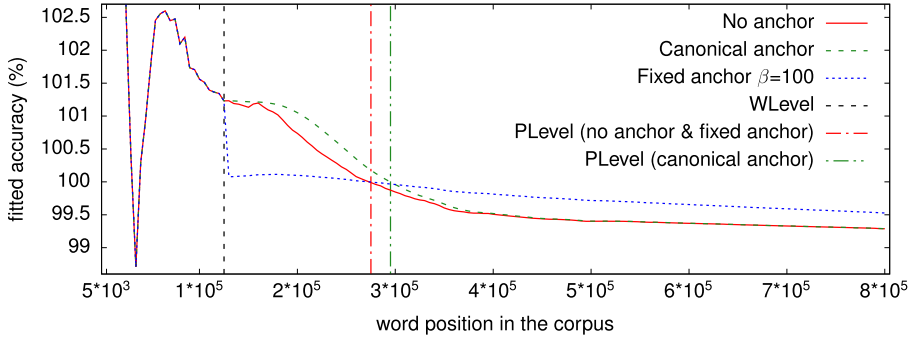
Fig. 3. Working and prediction levels without and with canonical anchors for fnTBL on Frown, with details in zoom.

These new learning trends differ from standard ones in the use of fitting points at infinity, while in practice they are located as far as the computer allows. The use of anchors to improve robustness responds to the idea that extra observations facilitate the realignment of the monotony for the asymptotic backbone, its residual at the point of infinity being the maximum degree of smoothing applicable at a given learning trend. This should be enough for small irregularities, thus limiting the strategy to levels after wlevel. A simple example is *canonical anchoring*.

Theorem 1. (Canonical anchoring) *Let $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ be a learning trace with asymptotic backbone $\{\alpha_i\}_{i\in\mathbb{N}}$ and $\{\hat{\mathcal{A}}_i(\infty)\}_{i>\omega}$ the sequence defined from its wlevel ω as*

$$\hat{\mathcal{A}}_{\omega+1}(\infty) := \alpha_\omega \qquad \hat{\mathcal{A}}_{i+1}(\infty) := \hat{\alpha}_i := \lim_{x\to\infty} \hat{\mathcal{A}}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}](x) \tag{7}$$

with $\hat{\mathcal{A}}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ a curve fitting $\{[x_j, \mathcal{A}_{\infty}[\mathcal{D}](x_j)], \ x_j := \|\mathcal{D}_j\|\}_{j=1}^i \cup [\infty, \hat{\mathcal{A}}_i(\infty)]$, $\forall i > \omega$. Then $\alpha_{\omega+i} \leq \hat{\alpha}_{\omega+i}$ (resp. $\alpha_{\omega+i} \geq \hat{\alpha}_{\omega+i}$), $\forall i \in \mathbb{N}$, when $\{\alpha_i\}_{i\in\mathbb{N}}$ is decreasing (resp. increasing). Also, $\{\hat{\mathcal{A}}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]\}_{i>\omega}$ is an anchoring learning trace of reference $[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \omega]$, with $\{\hat{\mathcal{A}}_i(\infty)\}_{i>\omega}$ its canonical anchors.

Proof. To see in [42]. □

Since in each cycle the anchor takes the value from the asymptote of the last learning trend, the technique described has a conservative nature, which translates into a slower convergence process. The effect of canonical anchoring in smoothing irregularities after the wlevel is illustrated vs. its absence, in our running example by a dashed line, in Fig. 3.

4.2. Correctness

It is addressed from the working hypotheses. That way, the uniform convergence of learning traces has been demonstrated, and the topology of the limit function described.

Theorem 2. *Let $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ be a learning trace with or without anchors. Then, its asymptotic backbone is monotonic and $\mathcal{A}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}] := \lim_{i\to\infty} {}^u \mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ exists, is positive definite, increasing, continuous and upper bounded by 100 in $(0, \infty)$.*

Proof. To see in [42]. □

This provides a way to estimate a learning curve $\mathcal{A}_{\infty}[\mathcal{D}]$ by iteratively approximating the function $\mathcal{A}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}]$, while a proximity criterion also needs to measure the convergence (resp. error) threshold at each stage. Namely, after fixing a level i in a learning trace $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$, we have to calculate an upper bound for the distance between $\mathcal{A}_j^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ and $\mathcal{A}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ (resp. $\mathcal{A}_{\infty}[\mathcal{D}]$) in the interval $[\|\mathcal{D}_j\|, \infty)$, $\forall j \geq i$.

A previous result is needed. Let $\{(q_{i,x}^{i-1}, q_{i,y}^{i-1})\}_{i\geq 4}$ (resp. $\{(p_{i,x}^{i-1}, p_{i,y}^{i-1})\}_{i\geq 4}$) be the sequence of the last (resp. first, if existing) points in $\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}] \cap \mathcal{A}_{i-1}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$. Then, $\{q_{i,x}^{i-1}\}_{i\geq 4}$ and $\{q_{i,y}^{i-1}\}_{i\geq 4}$ (resp. $\{p_{i,x}^{i-1}\}_{i\geq 4}$ and $\{p_{i,y}^{i-1}\}_{i\geq 4}$) are monotonic, except perhaps when there is a transition from one (resp. two) to two (resp. one) intersection points at a level ι, or when we introduce/modify anchors. In that case, $q_{\iota,y}^{i-1}$ and $q_{\iota+1,y}$ (resp. $p_{\iota,y}^{i-1}$ and $p_{\iota+1,y}$) may momentarily invert their relative positions, and the same applies to $q_{\iota,x}^{i-1}$ and $q_{\iota+1,x}$ (resp. $p_{\iota,x}^{i-1}$ and $p_{\iota+1,x}$). We then say that ι is a *level of rupture* for $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$. The order in $\mathbb{N}$ is also extended to $\mathbf{\overline{N}} := \mathbb{N} \cup \{\infty, \boldsymbol{\infty}\}$, in such a way that $\boldsymbol{\infty} > \infty > i > 0$, $\forall i \in \mathbb{N}$.
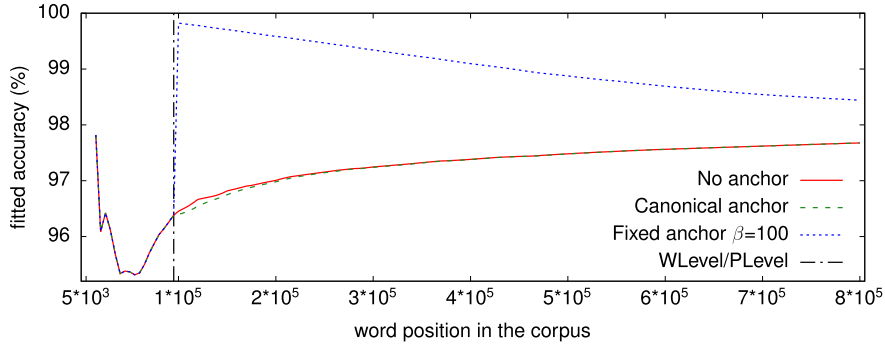
Fig. 4. Asymptotic backbones without and with fixed anchors for maxent on penn.

Theorem 3. (Correctness) *Let $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ be a learning trace with or without anchors, with wlevel ω, and $y = \alpha_i$ the asymptote for $\mathcal{A}_i^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$, $\forall i \in \mathbf{N} := \mathbb{N} \cup \{\infty, \boldsymbol{\infty}\}$. Let also $(q_{i,x}^{i-1}, q_{i,y}^{i-1})$ be the last point in $\mathcal{A}_i^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}] \cap \mathcal{A}_{i-1}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$, $\forall i \geq 4$, $i \neq \iota$, with ι level of rupture for $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$. We then have, using for the occasion the notation $\mathcal{A}_{\boldsymbol{\infty}}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ to refer $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}]$, that*

$$\mid [\mathcal{A}_k^{\pi} - \mathcal{A}_j^{\pi}][\mathcal{D}_{\sigma}^{\mathcal{K}}](x) \mid \leq \varepsilon_i := \mid q_{i,y}^{i-1} - \alpha_i \mid, \; \forall k, j \geq i \geq 4, \; \forall x \in [q_{i,x}^{i-1}, \infty) \tag{8}$$

$$(\text{resp. } \mid [\mathcal{A}_k^{\pi} - \mathcal{A}_j^{\pi}][\mathcal{D}_{\sigma}^{\mathcal{K}}](x) \mid \leq \varepsilon_i := \mid q_{\boldsymbol{\infty},y}^i - \alpha_{\boldsymbol{\infty}} \mid, \; \forall k, j \geq i \geq 1, \; \forall x \in [q_{\boldsymbol{\infty},x}^i, \infty)) \tag{9}$$

if $\{\alpha_i\}_{i>\omega+1}$ is decreasing (resp. *increasing*), *with $\{\varepsilon_i\}_{i>\omega+1}$ decreasing and convergent to 0*

Proof. To see in [42]. □

This result establishes the uniform convergence [1] of the learning trace $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ to the learning curve $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}_{\sigma}]$. In particular, this implies that the curve $\mathcal{A}_{\infty}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ we are iteratively approximating matches the latter if the training process is long enough and, therefore, $\alpha_{\infty} = \alpha_{\boldsymbol{\infty}}$. Sadly, the result only has a practical reading when the asymptotic backbone is decreasing, as it was in Fig. 2. Otherwise, the bound depends on the final accuracy we want to estimate ($\alpha_{\boldsymbol{\infty}}$), as with opennlp maxent (see *opennlp.apache.org/*) on the section of the *Wall Street Journal* (wsj) in the penn treebank [27]. In this case, the asymptotic backbone is increasing, as reflects the continuous line in Fig. 4. This gap must therefore be closed to guarantee the operability of the approach.

4.3. Completeness

The bulk of our work focuses on this issue, through research on anchoring as a tool to force the dynamics of convergence on learning traces and obtain a decreasing asymptotic backbone, thus ensuring the completeness sought. As a first step, we introduce a sufficient condition to identify anchors verifying such a property.

Theorem 4. *Let $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ be a learning trace with wlevel ω and plevel $\wp$, $y = \alpha_{\boldsymbol{\infty}}$ the asymptote for the learning curve $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}]$ and $\{\hat{\mathcal{A}}_i(\infty)\}_{i>\omega} \subset \mathbb{R}^+$ a convergent sequence such that:*

$$\hat{\mathcal{A}}_i(\infty) \geq \alpha_i, \; \forall i > \wp \tag{10}$$

$$\hat{\mathcal{A}}_i(\infty) - \hat{\mathcal{A}}_{i+1}(\infty) \geq \hat{\rho}_i(\infty) - \hat{\rho}_{i+1}(\infty), \; \forall i > \omega \tag{11}$$

Let also $\hat{\mathcal{A}}_i^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ be the learning trend with anchor $\hat{\mathcal{A}}_i(\infty)$ and asymptote $y = \hat{\alpha}_i$, and $\hat{\rho}_i(\infty) := \hat{\mathcal{A}}_i(\infty) - \hat{\alpha}_i$, $\forall i > \omega$. Then, $\{\hat{\mathcal{A}}_i^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]\}_{i>\omega}$ is an anchoring learning trace of reference $[\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \omega]$, such that $\{\hat{\alpha}_i\}_{i>\omega}$ is decreasing.

Proof. To see in Appendix A. □

So, a criterion to generate a learning trace with decreasing backbone is to select a set of anchors never below the accuracy extrapolated at each cycle to the total training data base (10), as long as the learner does not override the readjustment applied by the anchoring (11). Intuitively simple, we will first study the practical utility of this idea for the case of the previously introduced canonical anchors.

Theorem 5. *Let $\hat{\mathcal{A}}_i^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ be a learning trace with canonical anchoring. We then have that if the asymptotic backbone of the reference is decreasing, then the same thing applies to that of the former.*

Proof. To see in Appendix A. □

Unfortunately, this result does not settle the question at hand, i.e. to guarantee decreasing asymptotic backbones by using anchors in order to have a practical absolute measure of the convergence of learning traces. As shown above, when using a canonical approach, this is ensured only if the reference already verifies it. Otherwise, the resulting asymptotic backbone can also be increasing, as is shown by the dashed line in Fig. 4, and another anchoring strategy is needed to respond to the challenge.

4.3.1. Fixed anchoring

Learning trends are fitting curves on the set of observations available at that time and, when using anchoring, also the value of the latter associated to the point of infinity. This is the key to making the proximity criterion in Theorem 3 fully operational, because the sum total of residuals on such curves is null. So, to achieve decreasing asymptotic backbones it suffices to fix anchors with negative or null residual, which is to say that they must rise above all existing and future observations, for example using values higher or equal than the maximum accuracy (100).

Theorem 6. (Fixed anchoring) *Let $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ be a learning trace with WLevel ω and $\hat{\mathcal{A}}_{i}^{\beta \pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ the learning trend with anchor $\hat{\mathcal{A}}_{i}^{\beta}(\infty) := \beta \geq 100, \forall i > \omega$. Then, $\hat{\mathcal{A}}^{\beta \pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}] := \{\hat{\mathcal{A}}_{i}^{\beta \pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]\}_{i>\omega}$ is an anchoring learning trace of reference $[\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \omega]$ with asymptotic backbone decreasing. We call $\{\hat{\mathcal{A}}_{i}^{\beta}(\infty)\}_{i>\omega}$ the* fixed anchors *of value β for $\hat{\mathcal{A}}^{\beta \pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$.*

Proof. To see in Appendix A. □

Fixed anchoring therefore guarantees the hypotheses under which we can determine a computable estimation of the convergence and error thresholds in absolute terms. Namely, it allows us to generate learning traces with decreasing asymptotic backbones, regardless of the training process considered. An example of this is shown in Fig. 4, where the monotony of the starting asymptotic backbone changes from increasing to decreasing when using fixed anchors of value $\beta = 100$. In these conditions, the completeness of our abstract model derives immediately.

Theorem 7. (Completeness) *Let $\hat{\mathcal{A}}^{\beta \pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ be a learning trace of fixed anchoring with WLevel ω and $y = \alpha_i$ the asymptote for $\hat{\mathcal{A}}_{i}^{\beta}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \forall i \in \mathbb{\bar{N}} := \mathbb{N} \cup \{\infty, \infty\}$. Let also be $(q_{i,x}^{i-1}, q_{i,y}^{i-1})$ the last point in $\hat{\mathcal{A}}_{i}^{\beta}[\mathcal{D}_{\sigma}^{\mathcal{K}}] \cap \hat{\mathcal{A}}_{i-1}^{\beta}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \forall i \geq 4, i \neq \iota$, with ι level of rupture for $\hat{\mathcal{A}}^{\beta \pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$. We then have, using for the occasion the notation $\hat{\mathcal{A}}_{\infty}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ (resp. $\hat{\mathcal{A}}_{\infty}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$) to refer $\mathcal{A}_{\infty}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ (resp. $\mathcal{A}_{\infty}[\mathcal{D}]$), that*

$$| [\hat{\mathcal{A}}_{k}^{\beta} - \hat{\mathcal{A}}_{j}^{\beta}][\mathcal{D}_{\sigma}^{\mathcal{K}}](x) | \leq \varepsilon_i := | q_{i,y}^{i-1} - \alpha_i |, \forall k, j \geq i > \omega + 1, \forall x \in [q_{i,x}^{i-1}, \infty) \tag{12}$$

with $\{\varepsilon_i\}_{i>\omega+1}$ decreasing and convergent to 0. We call the smallest $\iota \geq 4$ for which $| q_{\iota,y}^{\iota-1} - \alpha_\iota | \leq \tau$, the threshold level *for $\tau \in \mathbb{R}^+$.*

Proof. To see in Appendix A. □

Contrary to what happened with canonical anchors, the fixed ones free us from checking the decrease in the asymptotic backbone. Following Theorem 3, this provides a practical and extremely simple criterion for implementing a proximity condition measuring absolute thresholds, henceforward referred to as $\mathcal{H}_a$.

More in detail, given a learning trace with fixed anchoring $\hat{\mathcal{A}}^{\beta \pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ and a value $\tau \in \mathbb{R}^+$, we can assure that, once the corresponding threshold level ι has been located:

$$| [\hat{\mathcal{A}}_{\infty}^{\beta} - \hat{\mathcal{A}}_{j}^{\beta}][\mathcal{D}_{\sigma}^{\mathcal{K}}](x) | \leq \varepsilon_\iota := | q_{\iota,y}^{\iota-1} - \alpha_\iota | \leq \tau, \forall j \geq \iota > \omega + 1, \forall x \in [q_{\iota,x}^{\iota-1}, \infty) \tag{13}$$

Namely, all estimates in the interval $[q_{\iota,x}^{\iota-1}, \infty) \supseteq [\|\mathcal{D}_\iota\|, \infty)$ for the learning trends computed from the ι level are at a distance from the curve $\mathcal{A}_{\infty}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ to which we converge (resp. the learning curve $\mathcal{A}_{\infty}[\mathcal{D}]$), which is less than the threshold τ set.

As for canonical anchors, the fixed ones also contribute to a litte delay in the convergence, as can be seen in Fig. 3 because their values are always higher than the asymptotes associated with the learning trends. One way to reduce this undesirable side effect is to provide the anchoring with mechanisms that allow it to adapt to the dynamics of the learning process.
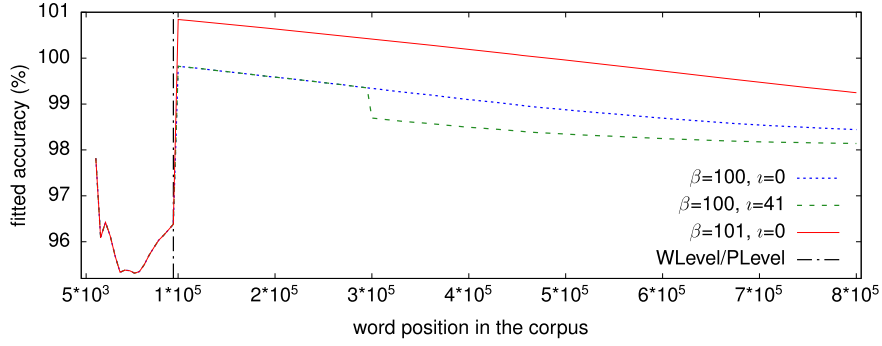
Fig. 5. Asymptotic backbones with fixed anchors for MAXENT on PENN (look-ahead ι, value β).

4.3.2. Endowing fixed anchoring with flexibility

The goal is to define a configurable family of anchorings ensuring the completeness of the proposal, thus allowing us to control the performance via an appropriate setting. Our starting point is the fixed anchor concept described above. Since the residuals at the point of infinity are then negative, we can fine tune anchors from the approximations for accuracy generated as learning progresses, without compromising our objective.

Theorem 8. (Fixed anchoring with look-ahead) *Let $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ be a learning trace with PLEVEL $\wp$, $\beta \geq 100$, $\ell \in \mathbb{N}$ and $\{\hat{\mathcal{A}}_i^{\beta,\ell}(\infty)\}_{i>\omega}$ the sequence defined from its WLEVEL ω by*

$$\hat{\mathcal{A}}_i^{\beta,\ell}(\infty) := \beta, \; \forall \, \wp + \ell + 1 > i > \omega \qquad \hat{\mathcal{A}}_i^{\beta,\ell}(\infty) := \hat{\alpha}_{\wp+\ell}^{\beta,\ell}, \; \forall i \geq \wp + \ell + 1 \qquad (14)$$

Let also $\hat{\mathcal{A}}_i^{\beta,\ell}[\mathcal{D}_\sigma^\mathcal{K}]$ be the learning trend with anchor $\hat{\mathcal{A}}_i^{\beta,\ell}(\infty), \forall i > \omega$. Then, $\hat{\mathcal{A}}^{\beta,\ell}_\pi[\mathcal{D}_\sigma^\mathcal{K}] := \{\hat{\mathcal{A}}_i^{\beta,\ell}[\mathcal{D}_\sigma^\mathcal{K}]\}_{i>\omega}$ is an anchoring learning trace of reference $[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \omega]$ and asymptotic backbone $\{\hat{\alpha}_i^{\beta,\ell}\}_{i>\omega}$ decreasing, and we call $\{\hat{\mathcal{A}}_i^{\beta,\ell}(\infty)\}_{i>\omega}$ its set of fixed anchors with look-ahead ℓ and value β.

Proof. To see in Appendix A. □

Intuitively, we are talking about a learning trace with conventional fixed anchoring, in which the anchor is subject to revision once the study of the levels in the interval $[\omega + 1, \wp + \ell]$ has been completed. Since $\omega \leq \wp$, this interval includes $[\wp + 1, \wp + \ell]$, which in either case allows us to take advantage of the knowledge provided by the first ℓ iterations from the PLEVEL $\wp$. As these are the best performing training cycles – together with the one associated with level $\wp$, in the case where $\omega = \wp$ – among those with real predictive capability, convergence can be expected to accelerate significantly once the anchor has been updated.

To illustrate this, we look again at the learning process shown in Fig. 4, to compare in Fig. 5 the asymptotic backbones associated to some value/look-ahead combinations, focusing on two use cases: different look-aheads ($\iota = 0$ and $\iota = 41$) with the same value ($\beta = 100$) and the same look-ahead ($\iota = 0$) with different values ($\beta = 100$ and $\beta = 101$). In the former scenario, we check how a non-trivial look-ahead ($\iota = 41$) causes the desired effect. Also, as might be expected, the second one suggests that the closer the anchor to the real accuracy, the faster the convergence. All the above underscores the importance of an in-depth study on the impact of values and look-aheads on accuracy prediction. The objective is to establish whether these first impressions have a formal basis that allows us to effectively categorize the anchoring strategies described.

Theorem 9. (Anchoring categorization) *Let $\hat{\mathcal{A}}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$, $\hat{\mathcal{A}}^{\eta,\iota}_\pi[\mathcal{D}_\sigma^\mathcal{K}]$, $\hat{\mathcal{A}}^{\beta,\iota}_\pi[\mathcal{D}_\sigma^\mathcal{K}]$ and $\hat{\mathcal{A}}^{\beta,J}_\pi[\mathcal{D}_\sigma^\mathcal{K}]$ be learning traces of reference $[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \omega]$, generated from canonical and fixed anchors with look-aheads $\iota, J \in \mathbb{N} \cup \{0\}$ for values $\eta, \beta \geq 100$ respectively. We then have that, in any case (resp. when $\{\alpha_i\}_{i \in \mathbb{N}}$ decreasing), it verifies that $\forall i > \omega$*

$$\mid \alpha_i - \alpha_{\boldsymbol{\infty}} \mid \leq \mid \hat{\alpha}_i - \alpha_{\boldsymbol{\infty}} \mid \qquad (15)$$

and also $\forall J > \iota, \eta > \beta, i \geq \wp + J$

$$\mid \hat{\alpha}_i^{\beta,J} - \alpha_{\boldsymbol{\infty}} \mid \leq \mid \hat{\alpha}_i^{\beta,\iota} - \alpha_{\boldsymbol{\infty}} \mid \leq \mid \hat{\alpha}_i^{\eta,\iota} - \alpha_{\boldsymbol{\infty}} \mid \qquad (16)$$

$$(resp. \; \mid \alpha_i - \alpha_{\boldsymbol{\infty}} \mid \leq \mid \hat{\alpha}_i - \alpha_{\boldsymbol{\infty}} \mid \leq \mid \hat{\alpha}_i^{\beta,J} - \alpha_{\boldsymbol{\infty}} \mid \leq \mid \hat{\alpha}_i^{\beta,\iota} - \alpha_{\boldsymbol{\infty}} \mid \leq \mid \hat{\alpha}_i^{\eta,\iota} - \alpha_{\boldsymbol{\infty}} \mid) \qquad (17)$$

with $\{\hat{\alpha}_i\}_{i>\omega}$, $\{\hat{\alpha}_i^{\eta,\iota}\}_{i>\omega}$, $\{\hat{\alpha}_i^{\beta,\iota}\}_{i>\omega}$, $\{\hat{\alpha}_i^{\beta,J}\}_{i>\omega}$ and $\{\alpha_i\}_{i \in \mathbb{N}}$ their corresponding asymptotic backbones, and $\alpha_{\boldsymbol{\infty}}$ the asymptotic value for the learning curve $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}]$.

Proof. To see in Appendix A. □

The result guides the choice of anchoring. So, the fastest way to converge when the working hypotheses verify is to avoid fixed anchors, except if the asymptotic backbone is not decreasing. When this is quasi-decreasing because only the testing hypotheses are guaranteed, the canonical strategy is the most adequate. Finally, fixed anchoring with look-ahead is the alternative when no data about the training are available. The convergence speed here is inversely proportional to the value, which is why the best option is 100, the minimum one. Once a value is selected, the look-ahead introduces an extra factor to speed up the convergence according to its length, but only from the time the anchor is updated. Because of this, our objective could be reached before the latter is activated, in such a way that a smaller look-ahead might be more effective. Namely, an optimal choice depends on the convergence threshold – that matches, by Theorem 3, the error one – we are trying to identify, thus suggesting an iterative approach for dealing with it. We then make the decision to depend on the degree of convergence reached with respect to that threshold, taking into account that the first reliable level for predictions is PLevel.

Definition 6. (Percentage of uncovered threshold) *Let* $\hat{\mathcal{A}}_{\hat{\pi}}^{\beta}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ *be a learning trace with fixed anchoring and* PLevel $\wp$*, and* τ *a threshold for a proximity condition* $\mathcal{H}$*. We define its* percentage of uncovered threshold for τ on $\mathcal{H}$ at a level $\ell > \wp + 1$ as

$$\text{PUT}[\hat{\mathcal{A}}_{\hat{\pi}}^{\beta}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \tau, \mathcal{H}](\ell) := \begin{cases} 100 * \frac{|\alpha_{\infty} - \hat{\alpha}_{\ell}^{\beta}|_{\mathcal{H}} - \tau}{|\alpha_{\infty} - \hat{\alpha}_{\wp+2}^{\beta}|_{\mathcal{H}} - \tau} & \text{if } |\alpha_{\infty} - \hat{\alpha}_{\ell}^{\beta}|_{\mathcal{H}} \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

with $|\alpha_{\infty} - \hat{\alpha}_{\ell}^{\beta}|_{\mathcal{H}}$ *the estimates of* $\mathcal{H}$ *for* $|\alpha_{\infty} - \hat{\alpha}_{\ell}^{\beta}|$*, and* $\hat{\alpha}_{\ell}^{\beta}$ (*resp.* α_{∞}) *the asymptotic value for the learning trend* (*resp. learning curve*) $\hat{\mathcal{A}}_{\ell}^{\beta}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ (*resp.* $\mathcal{A}_{\infty}[\mathcal{D}]$).

The PUT takes values in the interval $[0, 100]$ and is decreasing in the level covered. Its geometric interpretation is shown in Fig. 6. Its minimum is 0 and is reached when the estimated degree of approximation improves or equals the threshold τ set by the user, in asymptotic terms. That is, when the efficiency of the last asymptotic value as fixed anchor can no longer be improved, which implies that the current level is the last one at which a possible update of the fixed anchor based on learning asymptotic values makes sense. As for the maximum value, it is reached at level $\wp + 2$, the first at which the anchor could be updated and is 100, unless at that point the convergence is already effective.

Once we have captured the concept of PUT and the user has selected a particular value, we can assess for which particular level a fixed anchoring should be updated. That is, we are in a position to determine the look-ahead that matches the user's requirements.

Definition 7. (Minimal look-ahead for a tentative PUT value in fixed anchoring) *Let* $\hat{\mathcal{A}}_{\hat{\pi}}^{\beta}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$ *be a learning trace with fixed anchoring and* PLevel $\wp$*,* τ *a threshold for a proximity condition* $\mathcal{H}$ *and* $\zeta \in [0, 100]$ *the* PUT *value we want to reach before updating the anchor. We then define*

$$\lambda := \min_{\ell > \wp + 1} \{\text{PUT}[\hat{\mathcal{A}}_{\hat{\pi}}^{\beta}[\mathcal{D}_{\sigma}^{\mathcal{K}}, \tau, \mathcal{H}](\ell) \leq \zeta\} - \wp \tag{19}$$

as the minimal look-ahead for the tentative PUT value ζ.

Formally, the fact that PUT is monotonic and bound guarantees that the concept of minimal look-ahead is well-defined [1]. Regarding its geometric interpretation and calculation process, both of them are schematized in Fig. 6. We start from the asymptotic values α_{∞}, $\hat{\alpha}_{\ell}^{\beta}$ and $\hat{\alpha}_{\wp+2}^{\beta}$, which respectively correspond to the learning curve, the last learning trend and the first one for which the fixed anchor could be updated. We can then estimate, according the $\mathcal{H}$ criterion, the distance yet to be completed from the current level (resp. the first level at which a look-ahead would be applicable) to reach the required convergence threshold, by means of the value $|\alpha_{\infty} - \hat{\alpha}_{\ell}^{\beta}|_{\mathcal{H}} - \tau$ (resp. $|\alpha_{\infty} - \hat{\alpha}_{\wp+2}^{\beta}|_{\mathcal{H}} - \tau$).

By way of illustration, assuming that we want the update of a fixed anchoring to be activated once half of the convergence distance has been covered, it is sufficient to identify the first level $\ell > \wp + 1$ at which the PUT is less than or equal to 50 and then choose $\lambda := \ell - \wp$ as the minimal look-ahead for that tentative value. Returning to the example in Fig. 5 when the value of the fixed anchoring for the learning trace is $\beta = 100$, this results in a look-ahead of 41, whose associated asymptotic backbone shows the best results from among those represented.

5. The testing frame

The focus is now on evaluating the proposal, taking the generation of ML-based POS taggers as a case study. This first involves the design of a uniform framework, in the sense that its standards of evidence do not favour any particular *approximation technique*, namely any proximity condition nor particular setting. Once a learner and a training data base are fixed,
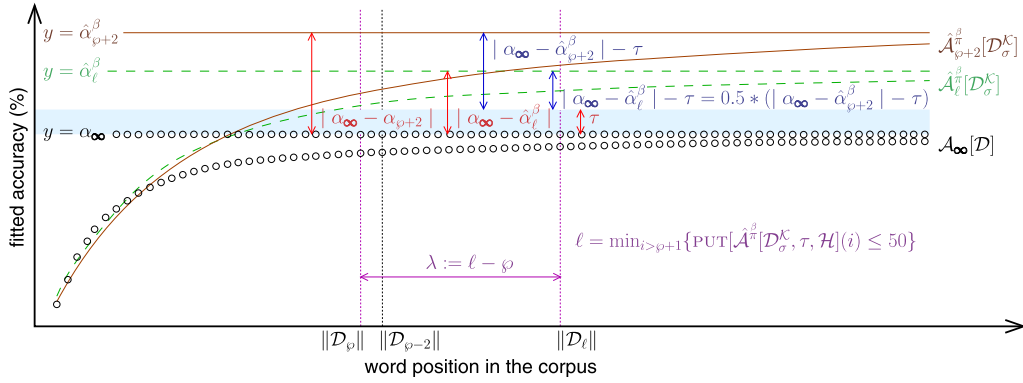
Fig. 6. Computing the minimal look-ahead λ for the tentative PUT value of 50 associated with a fixed anchoring.

the aim is to assess the impact of forcing the completeness of our argument, the balance between its costs and benefits, and also its stability against look-ahead variations.

To do this, we introduce the corresponding performance metric, together with its monitoring architecture for data collection. The latter captures the concept of testing round (*run*), which serves to normalize the conditions under which the experiments take place. Runs only distinguishable by their approximation technique are grouped around a baseline, in what we call a *local testing frame*, thus providing the environment we are looking for.

5.1. The monitoring structure

After setting an ML task represented by a learning trace $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$, the goal is to standardize the testing conditions, with a view to allowing for its objective assessment.

5.1.1. The testing rounds

Our evaluation basis is the *run*, a tuple $\hat{\mathcal{E}}^\mathcal{H} = [\hat{\mathcal{A}}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau, \mathcal{H}]$ characterized by a learning trace $\hat{\mathcal{A}}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ of reference $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ and anchoring $\hat{}$, a convergence or error threshold τ and a proximity condition $\mathcal{H}$. We can then express the PUT as a function on the runs, denoting is value on a given testing round $\hat{\mathcal{E}}^\mathcal{H}$ as PUT[$\hat{\mathcal{E}}^\mathcal{H}$], while the notion of *prediction level* is naturally extended as the one of its learning trace and denoting it by PLevel[$\hat{\mathcal{E}}^\mathcal{H}$]. A value CLevel[$\hat{\mathcal{E}}^\mathcal{H}$] is associated as its *convergence level*, the one from which $\mathcal{H}$ verifies for τ and the training stops. Given two runs $\hat{\mathcal{E}}^\mathcal{H} = [\hat{\mathcal{A}}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau, \mathcal{H}]$ and $\check{\mathcal{E}}^{\mathcal{H}'} = [\check{\mathcal{A}}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau, \mathcal{H}']$, they are *similar* when they are only distinguishable by the anchoring strategy used, i.e. when $\mathcal{H} = \mathcal{H}'$, and *dissimilar* otherwise. As our proposal requires decreasing asymptotic backbones, it is mandatory to use runs meeting such a condition to give a comprehensive understanding of the tests.

5.1.2. The testing scenarios

Our aim is to define run groupings to study the behaviour of an approximation technique beyond the qualitative considerations in Theorem 9. As the idea is to do it through a ratio with respect to a benchmark, it is necessary to compare runs sharing the reference but not the anchoring or the proximity condition. To this end, we introduce an order relation for these latter ones.

Theorem 10. *Let $\mathcal{L}$ be a set of runs only distinguishable by a proximity condition taken from Λ. Then, the relation*

$$\forall \mathcal{H}, \mathcal{H}' \in \Lambda, \ \mathcal{H} \succeq_\mathcal{L} \mathcal{H}' :\Leftrightarrow \text{CLevel}[\mathcal{E}^\mathcal{H}] \leq \text{CLevel}[\mathcal{E}^{\mathcal{H}'}], \forall \mathcal{E}^\mathcal{H}, \mathcal{E}^{\mathcal{H}'} \in \mathcal{L} \tag{20}$$

defines a partial order and we say that $\mathcal{H}$ is faster than $\mathcal{H}'$ on $\mathcal{L}$.

Proof. Trivial. □

Comparing runs also entails normalizing a threshold when it applies to proximity conditions with different scales, as with the absolute criterion just introduced and the relative one based on the layered correctness [42], which we refer to as $\mathcal{H}_a$ and $\mathcal{H}_r$, respectively. Once that happens, we first fix the relative threshold τ_r to be used with $\mathcal{H}_r$. The corresponding absolute one τ_a concerning $\mathcal{H}_a$ is then calculated, as Theorem 3 indicates for decreasing asymptotic backbones, from the level at which $\mathcal{H}_r$ determines the layered convergence. Such absolute thresholds are the ones referred to in the runs, which can then be grouped for testing purposes.

Definition 8. (Local testing frame) *Let $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ be a learning trace, τ a convergence or error threshold and Λ (resp. Γ) a set of proximity criteria (resp. anchoring strategies). We say that the collection*

$$\mathcal{L}_\Gamma^\Lambda[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau] := \{[\hat{\mathcal{A}}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau, \mathcal{H}], \ such\ that\ \hat{} \in \Gamma\ and\ \mathcal{H} \in \Lambda\} \tag{21}$$

is a local testing frame *iff exists* $\mathcal{F} \in \Lambda$ *which is the fastest on it.*

Intuitively, we are talking about sets of items only distinguishable by the anchoring and/or proximity condition, i.e. by the approximation technique considered. As $\emptyset \in \Gamma$, any local testing frame $\mathcal{L}_\Gamma^\Lambda[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau]$ includes the anchor-free runs $\mathcal{E}^\mathcal{H} :=$ $[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau, \mathcal{H}]$ whatever $\mathcal{H} \in \Lambda$. We baptize $\mathcal{E}^\mathcal{F} := [\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau, \mathcal{F}]$, the one using the fastest proximity criterion, as the *baseline run*. Since the anchors decelerate the convergence, their absence automatically increases its speed, depending on the proximity criteria used. That way, from a computational viewpoint, the baseline is the most efficient learning configuration in a local testing frame.

5.2. Performance metric

According to the principle of *maximum expected utility* (MEU) [28], we interpret the performance associated with a run as the search for a satisfactory cost/benefit trade-off. In that regard, any estimate of such performance requires the prior formalization of the concepts of cost and benefit for a run within its local testing frame, i.e. within its referential context. At this point, since we are interested in studying the behaviour of different anchorages and/or proximity conditions through a collection of local testing frames, it will be necessary to resort to measures relative to the baseline runs.

5.2.1. Cost of a run

The effort of convergence for a run identifies with its CLevel, provided it may be expressed in terms of the number of iterations needed to attain the degree of refinement required. We can do this by considering the same computational reference and threshold in all runs compared, as occurs within a local testing frame $\mathcal{L}_\Gamma^\Lambda[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau]$, assuming that the costs associated to the anchoring (resp. proximity condition) itself are comparable for all $\hat{} \in \Gamma$ (resp. $\mathcal{H} \in \Lambda$). This also provides a simple way for normalizing the cost associated with a run, taking the baseline as a benchmark.

Definition 9. (Relative cost) *Let* $\hat{\mathcal{E}}^\mathcal{H} \in \mathcal{L}_\Gamma^\Lambda[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau]$ *be a run in a local testing frame of baseline* $\mathcal{E}^\mathcal{F}$. *We define its* relative cost *as*

$$\text{RC}(\hat{\mathcal{E}}^\mathcal{H}) := \frac{\text{CLevel}[\mathcal{E}^\mathcal{H}]}{\text{CLevel}[\hat{\mathcal{E}}^\mathcal{F}]} \in [1, \infty) \tag{22}$$

The RC is positive and greater the greater the number of epochs needed to converge, i.e. the faster the limit curve is approximated the more its value is reduced, which allows Theorem 9 to be interpreted in terms of computational costs. That way, its minimum is 1 and is reached when the cost is that of the baseline, thus justifying our interest in RCs as close as possible to the unit. However, a low cost is not enough to conclude the advisability of using absolute thresholds against relative ones. Unless the specifications explicitly require one or the other, such a decision should be the consequence of balancing costs and benefits.

5.2.2. Performance of a run

Understood as the balance between benefit and cost, the performance of a run in the context of its local test frame is assimilable to the ratio between the degree of accuracy achieved and the relative cost accumulated during the learning process. With this aim in mind, we still have to formalize the concept of *accuracy*. If we refer to a convergence (resp. error) threshold τ, this should be higher the better the fit of the latter to the difference between the curve $\mathcal{A}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ to which we converge (resp. the learning curve $\mathcal{A}_\infty[\mathcal{D}]$) and the converging learning trend, which is used to provide an estimate at which the proximity condition $\mathcal{H}$ is verified, as indicated in Theorem 7.

Definition 10. (Convergence and error accuracy) *Let* $\hat{\mathcal{E}}^\mathcal{H} \in \mathcal{L}_\Gamma^\Lambda[\hat{\mathcal{A}}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau]$ *be a run with* CLevel ℓ. *We define its* convergence (*resp.* error) accuracy *as*

$$\text{A}^c(\hat{\mathcal{E}}^\mathcal{H})\ (resp.\ \text{A}^e(\hat{\mathcal{E}}^\mathcal{H})) := \begin{cases} 0 & if\ \exists\, i \geq \iota,\ \delta_\ell^c(i)\ (resp.\ \delta_\ell^e(i)) > \tau \\ 100 * \frac{\max_{i \geq \iota}\{|\delta_\ell^c(i)|\ (resp.\ |\delta_\ell^e(i)|)\}}{\tau} & otherwise \end{cases} \tag{23}$$

with ι *the threshold level for* τ, *and* $\delta_\ell^c(i)$ (*resp.* $\delta_\ell^e(i)) :=| [\hat{\mathcal{A}}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}]\ (resp.\ \mathcal{A}_\infty[\mathcal{D}]) - \hat{\mathcal{A}}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}]](\|\mathcal{D}_i\|) |$ *the divergence of* $\hat{\mathcal{A}}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ *with respect to* $\hat{\mathcal{A}}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ (*resp.* $\mathcal{A}_\infty[\mathcal{D}]$) *at level i.*

Thus defined as a percentage, the accuracy corresponds to the intuitive concept, which justifies our choice for the name of these metrics. Indeed, we calculate $\text{A}^c(\hat{\mathcal{E}}^\mathcal{H})$ (resp. $\text{A}^e(\hat{\mathcal{E}}^\mathcal{H})$) as the degree of precision achieved by run $\hat{\mathcal{E}}^\mathcal{H}$ at its CLevel in the estimation of the convergence (resp. error) threshold τ, starting from the iteration ι delimiting the interval for the completeness condition described in Theorem 7. That way, the value is zero when the threshold τ is not reached in that
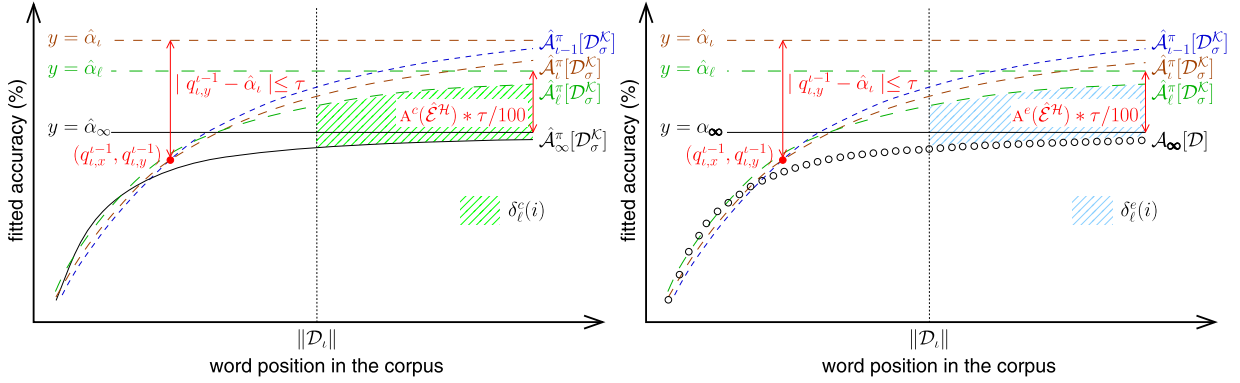
Fig. 7. Computing the convergence (resp. error) accuracy for a convergence (resp. error) threshold τ.

interval. For ease of understanding, the calculation process is illustrated in the left (resp. right) diagram of Fig. 7 for the convergence (resp. error) accuracy. It can be seen that the threshold level ι from which we search for the maximum value for the divergence δ_ℓ^c (resp. δ_ℓ^e) on the learning trend $\hat{\mathcal{A}}_\ell^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ associated with cLevel ℓ, and which in the case under consideration would be reached at its asymptote. The figure also shows the point $(q_{\iota,x}^{\iota-1}, q_{\iota,y}^{\iota-1})$, which marks the beginning of the domain of completeness for the threshold τ. However, to calculate this accuracy measure we need to know the curve $\hat{\mathcal{A}}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ (resp. $\mathcal{A}_\infty[\mathcal{D}]$). To address this issue, we assume a large enough set of observations provided by an omniscient oracle for the learning curve $\mathcal{A}_\infty[\mathcal{D}]$ through a sequence of contiguous individuals including the kernel $\mathcal{K}$. Henceforth, we refer to this set as the *horizon* of the learning trace $\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$. From this, $\hat{\mathcal{A}}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ (resp. $\mathcal{A}_\infty[\mathcal{D}]$) is estimated by the learning trend approximating such a set (resp. by such a set) of observations, together with its asymptotic value $\hat{\alpha}_\infty$ (resp. α_∞). Note that, following Theorem 3, $\alpha_\infty = \hat{\alpha}_\infty$. According to this, the calculation of error accuracy will be made assuming in each case the same asymptotic value for the set of observations as for the approximation considered of $\hat{\mathcal{A}}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}]$.

In practice, although we are interested in high values, it should be borne in mind that low accuracy does not necessarily imply a poor approximation process. Since the computationally more efficient runs are associated with larger convergence (resp. error) distances between epochs, it is more likely that it is in those runs where the divergences from the limit (resp. learning) curve will be smaller. In other words, runs with small RCs could eventually reach low accuracies, thus justifying the need for formalizing the concept of performance. As for accuracy, we differentiate between convergence and error performances to refer to the approximations of the limit and learning curves, respectively.

Definition 11. (Relative convergence and error performance) *Let* $\hat{\mathcal{E}}^\mathcal{H} \in \mathcal{L}_\Gamma^\Lambda[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau]$ *be. We define its* relative convergence (*resp.* relative error) performance *as*

$$\text{RP}^c(\hat{\mathcal{E}}^\mathcal{H}) \ (\text{resp. } \text{RP}^e(\hat{\mathcal{E}}^\mathcal{H})) := \frac{\text{A}^c(\hat{\mathcal{E}}^\mathcal{H}) \ (\text{resp. } \text{A}^e(\hat{\mathcal{E}}^\mathcal{H}))}{\text{RC}(\hat{\mathcal{E}}^\mathcal{H})} \in [0, 100] \tag{24}$$

Intuitively, the lower the RP in either of its interpretations, the more we could argue that an alternative approximation strategy should be considered. We are therefore interested in RPs as close as possible to 100.

6. The experiments

As said, they focus on learners for ML-based tagger generation, a demanding task in NLP. We thus need to introduce the linguistic resources and the testing space.

6.1. The linguistics resources

Corpora and POS tagger generators are selected from the most popular ones, as training data and learners respectively, the former together with their tag-sets:

1. The section of the WSJ in the PENN TREEBANK [27], of over 1,170,000 words.
2. The Freiburg-Brown (FROWN) of American English [18], of over 1,165,000 words.

PENN is annotated with POS tags and syntactic structures. By stripping it of the latter, it can be used to train POS taggers. To ensure well-balanced corpora, we have scrambled them at sentence level before testing.

We focus on models built from supervised learning, which make it possible to work with predefined tag-sets, thereby facilitating both the evaluation and the comprehension of the results:

1. In the category of stochastic methods and representing the *hidden Márkov models* (HMMs), we chose TNT [5]. We also include the TREETAGGER [34], which uses decision trees to generate the HMM, and MORFETTE [8], an averaged perceptron approach [10]. To illustrate the *maximum entropy models* (MEMs), MXPOST [31] and OPENNLP MAXENT [39]. Finally, the stanford POS tagger [39] combines features of HMMs and MEMs using a *conditional Márkov model*.

2. Under the heading of other approaches we consider FNTBL [29], an update of Brill [7], as example of transformation-based learning. As memory-based method we take the *memory-based tagger* (MBT) [11], while SVMTOOL [17] illustrates the behaviour with respect to support vector machines. We also use a perceptron-based training strategy with look-ahead, LAPOS [40].

This all ensures a good coverage of the linguistic resources for testing our proposal.

6.2. The testing space

We consider a collection $\mathcal{L}$ of local testing frames, with an entry for each combination of corpus and tagger. For each of these structures $\mathcal{L}_{\Gamma}^{\Lambda}[\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \tau]$, the size of the kernel $\mathcal{K}$ and the step function σ are fixed to $5 * 10^3$, while a power law family parameterizable by the trust region method [4] is chosen as accuracy pattern π. The WLevel of the runs is calculated from the values proposed in [42]: $\nu = 2 * 10^{-5}$, $\varsigma = 1$ and $\lambda = 5$. With respect to Γ, it includes both canonical and fixed anchoring. The proximity conditions are taken from $\Lambda = \{\mathcal{H}_a, \mathcal{H}_r\}$, as defined above. Since the asymptotic backbone of any of the runs concerned is decreasing, the applicability of $\mathcal{H}_a$ is guaranteed and these local testing frames are well-defined. Whatever the learning trace $\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}]$, we will consider as reference for testing purposes a horizon of 160 real observations provided by an omniscient oracle.

Having defined the testing structure, we address three aspects supporting the significance of the trials. The first relates to the exploitation of the training resources. Thus, as phrases are the smallest grammatical units with specific sense, samples should be aligned to the sentential distribution of the text. The second concerns the utility of the generated models, which depends on both the quality metrics being well-defined within the scope of the corpora and the reduction of variability phenomena. Finally, we tackle model optimization, i.e. the anchoring setting in each run.

6.2.1. Sampling fitting to sentence level and stability

We then need to adapt the learning schema. Given a corpus $\mathcal{D}$ with kernel $\mathcal{K}$ and a step function σ, we build the individuals $\{\mathcal{D}_i\}_{i \in \mathbb{N}}$ with $\mathcal{D}_i := \llbracket \mathcal{W}_i \rrbracket$ such that

$$\mathcal{W}_1 := \mathcal{K} \text{ and } \mathcal{W}_i := \mathcal{W}_{i-1} \cup \mathcal{I}_i, \ \mathcal{I}_i \subset \mathcal{D} \setminus \mathcal{W}_{i-1}, \ \|\mathcal{I}_i\| := \sigma(i), \ \forall i \geq 2 \tag{25}$$

where $\llbracket \mathcal{W}_i \rrbracket$ denotes the minimal set of sentences including the words in $\mathcal{W}_i$. This has no impact on our foundations and allows us to reap the maximum from training. Following [11,16], 10-fold cross validation confers stability on our measures.

6.2.2. Parameter tuning

As the speed of convergence relies on the anchoring used, fine-tuning is required to select a configuration close to the most efficient one and provide credibility to the tests when the strategy is parameterizable, which is what happens with fixed anchors. Given a local testing frame $\mathcal{L}_{\Gamma}^{\Lambda}[\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \tau]$, the way to do it is by choosing the optimal look-ahead for a value 100, namely the one with lowest RC. Moreover, as we will see, it is only necessary to focus on runs using the absolute proximity condition $\mathcal{H}_a$.

To this end, the potential look-aheads are studied in order of increasing size, which is the same as saying that they are explored according to their corresponding PUT, in decreasing order. For a uniform and complete monitoring of the procedure, its codomain [0,100] is covered with step 10. We then compare the RC of the runs using the minimal look-aheads corresponding to such a sequence of tentative PUTs.

Given ζ a tentative PUT for a run $[\mathcal{A}^{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \tau, \mathcal{H}_a]$, a candidate $[\hat{\mathcal{A}}^{100,\iota}_{\pi}[\mathcal{D}_{\sigma}^{\mathcal{K}}], \tau, \mathcal{H}_a]$ is generated from its minimal look-ahead ι and the RC calculated. The process is repeated for the next PUT until we locate the lowest RC. Hence, it is enough to identify the run associated to the turning point in such a RC sequence because, with respect to an increasing look-ahead, the performance is also increasing until reaching its maximum and then begins a decreasing trend. On this basis and to reduce the impact of irregularities, we chose that from which the window of increasing RCs is the largest one. So, it is hoped that the look-ahead is optimal with an error margin of 10% regarding the PUT metric.

6.3. The testing strategy

We do so according to the goals of our testing frame. That way, the cost of ensuring the completeness for $\mathcal{H}_a$ via fixed anchors is given by comparing within $\mathcal{L}$ the RC then applicable and the one estimated for similar runs when no anchoring or alternative canonical technique are used. These runs through $\mathcal{L}$ are hereafter referred to as $\mathcal{L}_{RC}$ and do not include the baseline ones.

To balance costs and benefits of $\mathcal{H}_a$ against $\mathcal{H}_r$, we contrast the RP in its two interpretations – RP^c and RP^e – when applying the former on a run in $\mathcal{L}$ with fixed anchors, and when no or canonical anchors are used for its dissimilar ones.

Table 1
RC monitoring in $\mathcal{L}_{RC}$.

		τ	No anchor + $\mathcal{H}_a$			Canonical + $\mathcal{H}_a$			Fixed + $\mathcal{H}_a$				
			PLevel	CLevel	RC	PLevel	CLevel	RC	PUT	ℓ_o	PLevel	CLevel	RC
Frown	fnTBL	1.50	55	58	1.00	59	77	1.33	89.38	6	55	100	1.72
	LAPOS	1.27	18	46	1.00	18	71	1.54	64.86	18	18	80	1.74
	MaxEnt	1.70	32	49	1.00	32	83	1.69	16.83	52	32	94	1.92
	MBT	1.95	43	51	1.00	51	86	1.69	29.40	43	43	98	1.92
	morfette	1.43	20	48	1.00	20	75	1.56	69.71	19	20	95	1.98
	MXPOST	2.84	22	30	1.00	22	31	1.03	75.63	11	22	59	1.97
	stanford	1.91	24	36	1.00	29	72	2.00	78.29	12	24	82	2.28
	SVMTool	1.41	41	52	1.00	46	89	1.71	76.02	10	41	92	1.77
	TnT	1.51	19	41	1.00	19	73	1.78	45.41	32	19	86	2.10
Penn	MBT	1.66	15	39	1.00	15	85	2.18	39.50	44	15	98	2.51
	MXPOST	1.40	17	28	1.00	17	50	1.79	18.73	37	17	57	2.04
	SVMTool	1.25	26	31	1.00	26	66	2.13	29.69	47	26	87	2.81

Table 2
RP^c monitoring in $\mathcal{L}_{RP}$.

		τ	No anchor + $\mathcal{H}_r$					Canonical + $\mathcal{H}_r$					Fixed + $\mathcal{H}_a$						
			PLevel	CLevel	A^c	RC	RP^c	PLevel	CLevel	A^c	RC	RP^c	PUT	ℓ_o	PLevel	CLevel	A^c	RC	RP^c
Frown	fnTBL	4.26	55	58	14.39	1.00	14.39	**59**	**60**	**16.42**	**1.03**	**15.88**	89.38	6	55	100	<u>4.24</u>	1.72	<u>2.46</u>
	LAPOS	3.47	18	46	6.15	1.00	6.15	**18**	**46**	**6.88**	**1.00**	**6.88**	64.86	18	18	80	8.38	1.74	4.82
	MaxEnt	4.42	32	49	4.03	1.00	4.03	**32**	**50**	**6.55**	**1.02**	**6.42**	16.83	52	32	94	<u>3.02</u>	1.92	<u>1.57</u>
	MBT	4.78	43	51	3.02	1.00	3.02	**51**	**52**	**10.66**	**1.02**	**10.45**	29.40	43	43	98	<u>1.11</u>	1.92	<u>0.58</u>
	morfette	3.67	20	48	6.02	1.00	6.02	**20**	**48**	**6.83**	**1.00**	**6.83**	69.71	19	20	95	6.98	1.98	3.53
	MXPOST	6.28	22	30	<u>1.81</u>	1.00	<u>1.81</u>	22	30	<u>0.85</u>	1.00	<u>0.85</u>	75.63	11	22	59	6.73	1.97	3.42
	stanford	4.31	24	36	7.88	1.00	7.88	**29**	**38**	**16.01**	**1.06**	**15.16**	78.29	12	24	82	5.33	2.28	2.34
	SVMTool	3.82	41	52	13.89	1.00	13.89	**46**	**53**	**16.50**	**1.02**	**16.19**	76.02	10	41	92	7.09	1.77	4.01
	TnT	3.69	19	41	5.45	1.00	5.45	**19**	**43**	**9.65**	**1.05**	**9.20**	45.41	32	19	86	5.77	2.10	2.75
Penn	MBT	3.72	15	39	18.89	1.00	18.89	15	39	<u>5.96</u>	1.00	<u>5.96</u>	39.50	44	15	98	12.85	2.51	5.11
	MXPOST	3.34	17	28	<u>2.22</u>	1.00	<u>2.22</u>	17	29	**4.40**	1.04	<u>4.25</u>	18.73	37	17	57	**14.70**	**2.04**	**7.22**
	SVMTool	2.64	26	31	23.09	1.00	23.09	**26**	**35**	**28.46**	**1.13**	**25.20**	29.69	47	26	87	22.94	2.81	8.18

Table 3
RP^e monitoring in $\mathcal{L}_{RP}$.

		τ	No anchor + $\mathcal{H}_r$					Canonical + $\mathcal{H}_r$					Fixed + $\mathcal{H}_a$						
			PLevel	CLevel	A^e	RC	RP^e	PLevel	CLevel	A^e	RC	RP^e	PUT	ℓ_o	PLevel	CLevel	A^e	RC	RP^e
Frown	fnTBL	4.26	55	58	14.39	1.00	14.39	**59**	**60**	**16.42**	**1.03**	**15.88**	89.38	6	55	100	<u>5.09</u>	1.72	<u>2.95</u>
	LAPOS	3.47	18	46	6.15	1.00	6.15	**18**	**46**	**6.88**	**1.00**	**6.88**	64.86	18	18	80	<u>5.23</u>	1.74	<u>2.73</u>
	MaxEnt	4.42	32	49	4.03	1.00	4.03	**32**	**50**	**6.55**	**1.02**	**6.42**	16.83	52	32	94	<u>5.23</u>	1.92	<u>2.73</u>
	MBT	4.78	43	51	3.02	1.00	3.02	**51**	**52**	**10.66**	**1.02**	**10.45**	29.40	43	43	98	<u>1.46</u>	1.92	<u>0.76</u>
	morfette	3.67	20	48	6.02	1.00	6.02	**20**	**48**	**6.83**	**1.00**	**6.83**	69.71	19	20	95	6.98	1.98	3.53
	MXPOST	6.28	22	30	<u>1.91</u>	1.00	<u>1.91</u>	22	30	<u>2.62</u>	1.00	<u>2.62</u>	75.63	11	**22**	**59**	**6.73**	**1.97**	**3.42**
	stanford	4.31	24	36	7.88	1.00	7.88	**29**	**38**	**16.01**	**1.06**	**15.16**	78.29	12	24	82	5.33	2.28	2.34
	SVMTool	3.82	41	52	13.89	1.00	13.89	**46**	**53**	**16.50**	**1.02**	**16.19**	76.02	10	41	92	7.09	1.77	4.01
	TnT	3.69	19	41	5.45	1.00	5.45	**19**	**43**	**9.65**	**1.05**	**9.20**	45.41	32	19	86	5.77	2.10	2.75
Penn	MBT	3.72	15	39	18.89	1.00	18.89	15	39	<u>7.80</u>	1.00	<u>7.80</u>	39.50	44	15	98	12.85	2.51	5.11
	MXPOST	3.34	17	28	<u>5.05</u>	1.00	<u>5.05</u>	17	29	<u>4.95</u>	1.04	<u>4.78</u>	18.73	37	17	57	**14.70**	**2.04**	**7.22**
	SVMTool	2.64	26	31	23.09	1.00	23.09	**26**	**35**	**28.46**	**1.13**	**25.20**	29.69	47	26	87	22.94	2.81	8.18

We exclude the dissimilar run with fixed anchor because Theorem 9 makes it clear that the only practical interest of this anchoring technique is to provide completeness for $\mathcal{H}_a$. Such runs through $\mathcal{L}$ are hereafter referred to as $\mathcal{L}_{RP}$, including the baseline ones.

Finally, in order to assess the stability of using $\mathcal{H}_a$ via fixed anchoring against look-ahead variations, we shall simply extend the set Γ of anchorings in each local testing frame $\mathcal{L}_\Gamma^\Lambda[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau]$, to include a selection of representative look-aheads. The runs involved in this study through $\mathcal{L}$ will be referred to as $\mathcal{L}_{RP_\ell}$.

6.4. The analysis of the results

The detail of the monitoring is compiled separately for $\mathcal{L}_{RC}$ (resp. $\mathcal{L}_{RP}$ and $\mathcal{L}_{RP_\ell}$) in Table 1 (resp. Tables 2–3 and 4–5) along with its PLevel. We also include the CLevel for each run, as well as its RC (resp. A^c and RP^c, and A^e and RP^e) value, which is better the closer it comes to 1 (resp. to 100). When using a fixed anchoring with value $\beta = 100$, the look-ahead is also visualized, together with the associated PUT in the case of the optimal value ℓ_o. These values are expressed to two
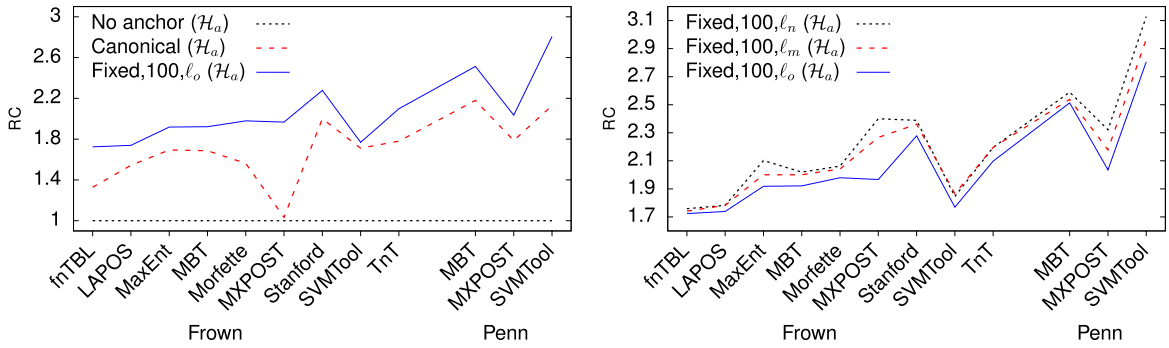
Fig. 8. RC values for runs in $\mathcal{L}_{\text{RC}}$ and $\mathcal{L}_{\text{RP}_\ell}$ (look-ahead ι, value $\beta = 100$).
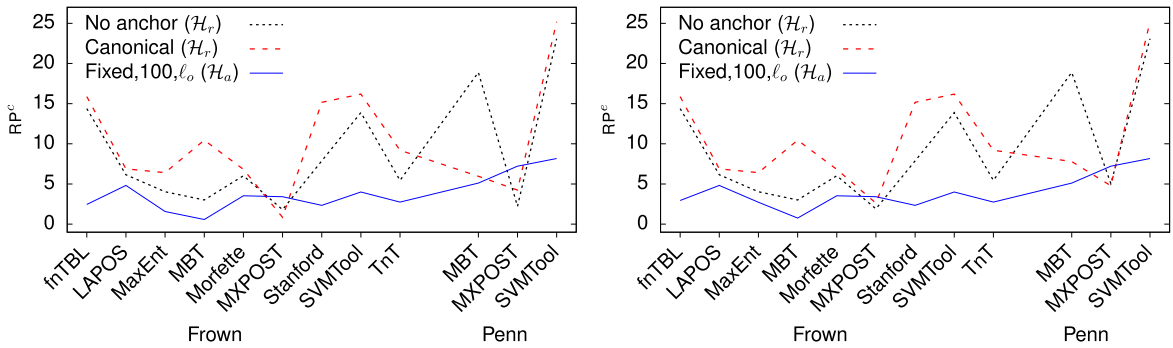

Fig. 9. RP^c and RP^e values for runs in $\mathcal{L}_{\text{RP}}$.

decimal places because of space limitations, using bold (resp. italic) fonts to mark the best results among all the (resp. baseline) runs, while all the calculations have been done to six decimal places of precision. Absent from these tables are the local testing frames whose baselines show increasing asymptotic backbones, as with fnTBL, LAPOS, MaxEnt, Morfette, Stanford, TnT and TreeTagger trained on Penn. We also discard those with any run converging beyond the boundaries of the training corpora, as with TreeTagger on Frown.

6.4.1. Evaluating the cost of using $\mathcal{H}_a$ via fixed anchoring

The benchmark measure is now RC, whose values are compiled in Table 1 and illustrated in the left-most diagram of Fig. 8 for the set $\mathcal{L}_{\text{RC}}$ of significant runs for this issue. Note that the sole aim of the estimates for the canonical technique is to illustrate the smaller impact of using $\mathcal{H}_a$ with a less invasive anchoring where possible, because in no way does the latter guarantee the completeness of such a proximity criterion.

In greater detail, RCs range from 1 for the baselines – anchor-free runs – to 2.81 for svMtool on Penn if fixed anchors are considered. In percentages, 77.78% of these values are less than 2, in an interval $[1, \infty)$ of possible costs. Analyzing each anchoring approach, this ratio grows to 100% for the baselines, dropping to 75% for those applying a canonical one and to 58.33% for the fixed strategy with optimal look-ahead. The best score is for anchor-free runs in all local testing frames, while canonical anchors always provide the second best. Taking into account that convergence speed and relative cost are proportional, these results exemplify Theorem 9. Specifically, they support both the greater computational complexity of applying fixed anchors and the superiority of anchor-free runs when the impact of irregularities on the learning process is limited.

6.4.2. Evaluating the costs and benefits of using $\mathcal{H}_a$ against $\mathcal{H}_r$

Our metric is now RP^c (resp. RP^e), whose values are compiled in Table 2 (resp. Table 3) and illustrated in the left-most (resp. right-most) diagram of Fig. 9 for the set $\mathcal{L}_{\text{RP}}$ of significant runs for this issue, as regards the treatment of convergence (resp. error) thresholds. Note that most of these two tables are identical, just 22.2% of the runs have different values, which we have underlined so they can be easily distinguished. The origin of this behaviour is the matching in most of the runs of the values for A^c and A^e. This, in turn, is a consequence of the fact that when the impact of irregularities in the learning process is limited, the maximum divergence values considered for their calculation often occur at the asymptotic level and, therefore, coincide.

In greater detail, anchor-free runs with $\mathcal{H}_r$ present the (resp. second) best RP^c and RP^e values in only one (resp. in ten) local testing frames, which corresponds to MBT on Penn. Regarding the use of fixed anchors with $\mathcal{H}_a$, it turns out to be the (resp. second) best choice in two (resp. in no) cases, MXPOST in both Frown and Penn. In all other local testing frames, the
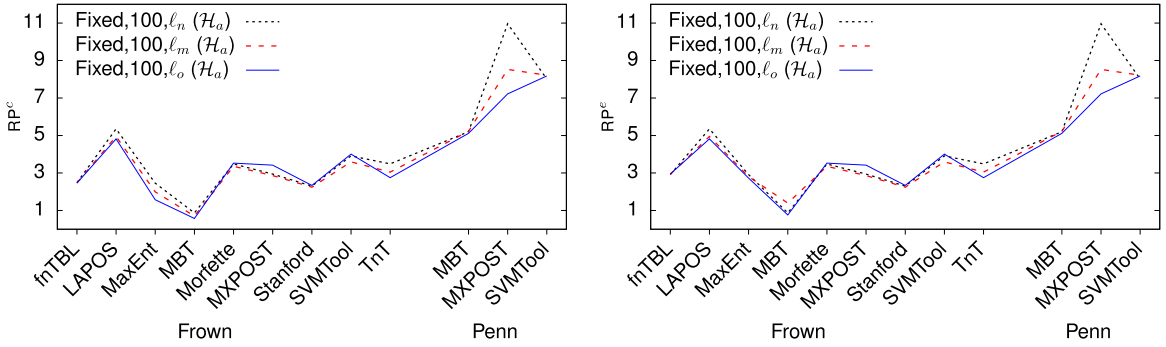
Fig. 10. RP^c and RP^e values for runs in $\mathcal{L}_{RP_\ell}$ (look-ahead ι, value $\beta = 100$).

Table 4
RP^c monitoring in $\mathcal{L}_{RP_\ell}$.

		τ	Fixed + $\mathcal{H}_a$						Fixed + $\mathcal{H}_a$						Fixed + $\mathcal{H}_a$					
			PUT	ℓ_o	PLevel	CLevel	A^c	RC	RP^c	ℓ_m	PLevel	CLevel	A^c	RC	RP^c	ℓ_n	PLevel	CLevel	A^c	RC	RP^c
Frown	fnTBL	4.26	89.38	6	55	100	4.24	1.72	2.46	3	55	101	4.31	1.74	2.47	0	55	102	4.37	1.76	2.49
	LAPOS	3.47	64.86	18	18	80	8.38	1.74	4.82	9	18	82	8.83	1.78	4.95	0	18	82	9.55	1.78	5.36
	MaxEnt	4.42	16.83	52	32	94	3.02	1.92	1.57	26	32	98	4.00	2.00	2.00	0	32	103	5.17	2.10	2.46
	MBT	4.78	29.40	43	43	98	1.11	1.92	0.58	22	43	102	1.40	2.00	0.70	0	43	103	1.76	2.02	0.87
	Morfette	3.67	69.71	19	20	95	6.98	1.98	3.53	10	20	98	6.85	2.04	3.35	0	20	99	7.18	2.06	3.48
	MXPOST	6.28	75.63	11	22	59	6.73	1.97	3.42	6	22	68	6.49	2.27	2.86	0	22	72	7.07	2.40	2.95
	stanford	4.31	78.29	12	24	82	5.33	2.28	2.34	6	24	85	5.30	2.36	2.24	0	24	86	5.50	2.39	2.30
	SVMTool	3.82	76.02	10	41	92	7.09	1.77	4.01	5	41	97	6.71	1.87	3.60	0	41	96	7.22	1.85	3.91
	TnT	3.69	45.41	32	19	86	5.77	2.10	2.75	16	19	90	6.69	2.20	3.05	0	19	90	7.65	2.20	3.49
Penn	MBT	3.72	39.50	44	15	98	12.85	2.51	5.11	22	15	99	13.26	2.54	5.22	0	15	101	13.47	2.59	5.20
	MXPOST	3.34	18.73	37	17	57	14.70	2.04	7.22	19	17	61	18.58	2.18	8.53	0	17	65	25.46	2.32	10.97
	SVMTool	2.64	29.69	47	26	87	22.94	2.81	8.18	24	26	92	24.38	2.97	8.21	0	26	97	25.10	3.13	8.02

Table 5
RP^e monitoring in $\mathcal{L}_{RP_\ell}$.

		τ	Fixed + $\mathcal{H}_a$						Fixed + $\mathcal{H}_a$						Fixed + $\mathcal{H}_a$					
			PUT	ℓ_o	PLevel	CLevel	A^c	RC	RP^c	ℓ_m	PLevel	CLevel	A^c	RC	RP^c	ℓ_n	PLevel	CLevel	A^c	RC	RP^c
Frown	fnTBL	4.26	89.38	6	55	100	5.09	1.72	2.95	3	55	101	5.10	1.74	2.93	0	55	102	5.13	1.76	2.92
	LAPOS	3.47	64.86	18	18	80	8.38	1.74	4.82	9	18	82	8.83	1.78	4.95	0	18	82	9.55	1.78	5.36
	MaxEnt	4.42	16.83	52	32	94	5.23	1.92	2.73	26	32	98	5.62	2.00	2.81	0	32	103	6.13	2.10	2.92
	MBT	4.78	29.40	43	43	98	1.46	1.92	0.76	22	43	102	2.79	2.00	1.40	0	43	103	1.76	2.02	0.87
	Morfette	3.67	69.71	19	20	95	6.98	1.98	3.53	10	20	98	6.85	2.04	3.35	0	20	99	7.18	2.06	3.48
	MXPOST	6.28	75.63	11	22	59	6.73	1.97	3.42	6	22	68	6.49	2.27	2.86	0	22	72	7.07	2.40	2.95
	stanford	4.31	78.29	12	24	82	5.33	2.28	2.34	6	24	85	5.30	2.36	2.24	0	24	86	5.50	2.39	2.30
	SVMTool	3.82	76.02	10	41	92	7.09	1.77	4.01	5	41	97	6.71	1.87	3.60	0	41	96	7.22	1.85	3.91
	TnT	3.69	45.41	32	19	86	5.77	2.10	2.75	16	19	90	6.69	2.20	3.05	0	19	90	7.65	2.20	3.49
Penn	MBT	3.72	39.50	44	15	98	12.85	2.51	5.11	22	15	99	13.26	2.54	5.22	0	15	101	13.47	2.59	5.20
	MXPOST	3.34	18.73	37	17	57	14.70	2.04	7.22	19	17	61	18.58	2.18	8.53	0	17	65	25.46	2.32	10.97
	SVMTool	2.64	29.69	47	26	87	22.94	2.81	8.18	24	26	92	24.38	2.97	8.21	0	26	97	25.10	3.13	8.02

canonical anchoring with $\mathcal{H}_r$ achieves the (resp. second) best results. Overall, as one would expect from its ability to adapt dynamically to the evolution of the learning process, the best performances come from the use of canonical anchors on $\mathcal{H}_r$. On the other hand, the strong static control imposed by the fixed ones to ensure the completeness for $\mathcal{H}_a$ should relegate their use to situations where we have no information about the magnitude of the irregularities and the monotonicity – increasing or decreasing – of the learning trace involved, or we simply need to ensure an absolute threshold.

6.4.3. Stability of using $\mathcal{H}_a$ via fixed anchoring against look-ahead variations

Although the results above illustrate the application of proximity conditions based on absolute thresholds, they were obtained from runs with fixed anchors and an optimal look-ahead ℓ_o resulting from a tuning process. The aim therefore is now to determine the real impact of such a process on run performance. In this context, we extend the set Γ of anchorings in each local testing frame $\mathcal{L}_\Gamma^\Lambda[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \tau]$ with the fixed ones of value $\beta = 100$ and null look-ahead ℓ_n, and also $\ell_m := \ell_o/2$. As ℓ_o is assumed to provide the best convergence speed for the proximity condition $\mathcal{H}_a$ in an increasing setting sequence, ℓ_n should supply the lowest one and ℓ_m intermediate ones. The monitoring of set $\mathcal{L}_\ell$, which brings the significant runs for this issue, is compiled (resp. illustrated) respectively for RP^c and RP^e in Tables 4 (resp. in the left-most diagram of Fig. 10) and 5 (resp. in the right-most diagram of Fig. 10). These tables also include the RC scores, which are illustrated separately in the right-most diagram of Fig. 8, and most of their entries are identical as for runs in $\mathcal{L}_{RP}$. Specifically, only 22.2% of the RP^c and RP^e values are different, and we have again used underlined text to highlight them.

In short, we find that RCs range from 1.72 for FNTBL on FROWN with a look-ahead ℓ_o, to 3.13 for SVMTOOL on PENN with ℓ_n. In percentages, 36.11% of these values are less than 2, in an interval $[1, \infty)$ of possible costs. Analyzing each anchoring approach, this ratio grows to 58.33% when using ℓ_o look-aheads and drops to 25% otherwise. The best score is for ℓ_o in all local testing frames, while ℓ_m provides the second best result in all cases but one, and ℓ_n in three. Note that ℓ_m and ℓ_n obtain the same RCs for LAPOS and TNT in FROWN. This again exemplifies the conclusions of Theorem 9 about anchor categorization, this time regarding the use of look-aheads, but also illustrates the usefulness and validity of the concept of minimal look-ahead for a tentative PUT as a mechanism for optimizing fixed anchorings.

As for RP^c (resp. RP^e), the optimal look-ahead ℓ_o gives the best values in four (resp. five) cases, while ℓ_m and ℓ_n do so in two (resp. three) and six (resp. four) ones, respectively. Regarding the second best choice, it corresponds to ℓ_o in only one (resp. one) case and in six (resp. five) ones to ℓ_m, while anchor-free runs reach that position five (resp. six) times. Overall, the best performances seem to correspond to the absence of look-ahead, followed by those associated with the use of the mean value ℓ_m and the optimum one ℓ_o, although the differences are very small as can be seen in Fig. 10. This suggests that the choice of look-ahead has a minor impact on the performance, thus allowing its use to be simplified, as it would no longer require a prior tuning phase.

7. Conclusions

We have responded to the challenge of estimating absolute convergence thresholds associated with the prediction of learning processes as a means of reducing training effort and the need for resources in the generation of ML-based systems. The goal is to get the most from a non-active adaptive sampling scheme used for that purpose, by limiting its application in time to what is strictly necessary, while avoiding the limitations of relative measures.

Our proposal proves its correctness with respect to its working hypotheses. Namely, it determines the cycle from which we can ensure that the threshold fixed by the user has been reached. Since this can only be established in practice when the successive estimates for accuracy are decreasing, the completeness of the technique is also stated. To do so, we demonstrate that it is possible to redirect the training dynamics in such a way that such a property verifies. This is achieved by properly using the concept of anchor, a conservative assessment of the final accuracy achievable by the learner, which is calculated from a sufficiently representative sample interpreted as an observation at the point of infinity for the calculation of the next estimate. Furthermore, since the primary function of anchoring is to compensate the irregularities in the learning process due to deviations in the working hypotheses, the proposal shows a good degree of robustness.

To reduce the slowdown caused on the pace of convergence by the use of anchors, we introduce a parameterizable family of these structures, categorizing them with respect to both the costs and the balance between these and their benefits. The tests, taking the generation of POS taggers in NLP as case study, corroborate our expectations. In particular, although the consideration of absolute thresholds applying our proposal entails greater computational cost, it has demonstrated its reliability and practical applicability, providing a stable and robust way to proceed when relative estimations of the learning curve are not sufficient for the development of ML applications.

CRediT authorship contribution statement

Manuel Vilares Ferro: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Víctor M. Darriba Bilbao:** Data curation, Software, Writing – review & editing. **Jesús Vilares Ferro:** Investigation, Validation, Visualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A. Proofs for Subsection 4.3

A.1. Proof of Theorem 4

Since $\hat{\rho}_i(\infty) := \hat{\mathcal{A}}_i(\infty) - \hat{\alpha}_i, \forall i > \omega$, the asymptotic backbone $\{\hat{\alpha}_i\}_{i>\omega}$ is decreasing iff it verifies that

$$\forall i > \omega, \ \hat{\alpha}_i \geq \hat{\alpha}_{i+1} \Leftrightarrow \hat{\mathcal{A}}_i(\infty) - \hat{\mathcal{A}}_{i+1}(\infty) \geq \hat{\rho}_i(\infty) - \hat{\rho}_{i+1}(\infty), \tag{26}$$

which derives immediately from Equation (11). To now complete the proof, we only need to demonstrate that $\{\hat{\alpha}_i\}_{i>\omega}$ converges to α_∞. As $\hat{\mathcal{A}}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ is a fitting curve for the values

$$\{[x_j, \mathcal{A}_\infty[\mathcal{D}](x_j)], \; x_j := \|\mathcal{D}_j\|\}_{j=1}^i \cup \{[\infty, \hat{\mathcal{A}}_i(\infty)]\}, \; \forall i > \omega \tag{27}$$

we then have that

$$\hat{\mathcal{A}}_i(\infty) \geq \hat{\alpha}_i := \lim_{x \to \infty} \hat{\mathcal{A}}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}](x), \; \forall i > \omega \tag{28}$$

If we also take into account that $\mathcal{A}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ is a curve fitting the set

$$\{[x_j, \mathcal{A}_\infty[\mathcal{D}](x_j)], \; x_j := \|\mathcal{D}_j\|\}_{j=1}^i \cup \{[\infty, \alpha_i]\}, \; \forall i > \omega \tag{29}$$

with anchors verifying the Equation (10), we then have that

$$\hat{\mathcal{A}}_i(\infty) \geq \hat{\alpha}_i := \lim_{x \to \infty} \hat{\mathcal{A}}_i^\pi[\mathcal{D}_\sigma^\mathcal{K}](x) \geq \alpha_i, \; \forall i > \wp \tag{30}$$

from which

$$\lim_{i \to \infty} \hat{\mathcal{A}}_i(\infty) \geq \lim_{i \to \infty} \hat{\alpha}_i \geq \lim_{i \to \infty} \alpha_i = \alpha_\infty \tag{31}$$

Moreover, the impact of the singularity $\hat{\mathcal{A}}_\infty^\pi[\mathcal{D}_\sigma^\mathcal{K}](\infty)$ in the generation of the learning trends $\{\hat{\mathcal{A}}_i[\mathcal{D}_\sigma^\mathcal{K}]\}_{i>\omega}$ decreases as the level ascends. Namely, α_∞ is a supremum for $\{\hat{\alpha}_i\}_{i>\omega}$ and $\lim\limits_{i \to \infty} \hat{\alpha}_i = \lim\limits_{i \to \infty} \alpha_i = \alpha_\infty$, which proves the thesis.

A.2. Proof of Theorem 5

Let $\{\alpha_i\}_{i\in\mathbb{N}}$ be the backbone for the reference $[\mathcal{A}^\pi[\mathcal{D}_\sigma^\mathcal{K}], \omega]$ of $\hat{\mathcal{A}}^\pi[\mathcal{D}_\sigma^\mathcal{K}]$ and $\wp$ the PLevel of the latter. By Theorem 4, it is enough to prove that its hypotheses verify. Focusing on Equation (10), let us first assume $i = \omega + 1$. Since by hypothesis $\{\alpha_i\}_{i\in\mathbb{N}}$ is decreasing, we conclude that

$$\hat{\mathcal{A}}_{\omega+1}(\infty) := \alpha_\omega \geq \alpha_{\omega+1} \tag{32}$$

Let us now assume that $i > \omega + 1$, as $\{\alpha_i\}_{i\in\mathbb{N}}$ is decreasing, Theorem 1 proves that $\hat{\alpha}_{\omega+i} \geq \alpha_{\omega+i}, \forall i > \omega$, from which we derive that

$$\hat{\mathcal{A}}_{i+1}(\infty) := \hat{\alpha}_i \geq \alpha_i \geq \alpha_{i+1} \tag{33}$$

and we can then affirm that $\hat{\mathcal{A}}_i(\infty) \geq \alpha_i, \; \forall i > \wp \geq \omega$, completing the proof in this case.

Regarding the compliance of the condition referred in Equation (11), as $\hat{\rho}_i(\infty) := \hat{\mathcal{A}}_i(\infty) - \hat{\alpha}_i$, demonstrating the inequality

$$\hat{\mathcal{A}}_i(\infty) - \hat{\mathcal{A}}_{i+1}(\infty) \geq \hat{\rho}_i(\infty) - \hat{\rho}_{i+1}(\infty), \; \forall i > 0 + \omega = \omega \tag{34}$$

which is to prove $\hat{\alpha}_i \geq \hat{\alpha}_{i+1}, \forall i > \omega$. As this was stated in Theorem 1, we conclude the thesis.

A.3. Proof of Theorem 6

By Theorem 4, as $\{\hat{\mathcal{A}}_i^\beta(\infty)\}_{i>\omega}$ is convergent to β, it is enough to prove that its hypotheses verify. To this end, the condition in Equation (10) becomes trivial because the PLevel $\wp$ is the first level after the working one ω from which the asymptotic backbone is below 100, and therefore

$$\hat{\mathcal{A}}_i^\beta(\infty) := \beta \geq 100 \geq \alpha_i, \; \forall i > \wp \tag{35}$$

With respect to the condition in Equation (11), given that $\hat{\mathcal{A}}_i^{\frac{\beta}{\pi}}[\mathcal{D}_\sigma^\mathcal{K}]$ is a curve fitting

$$\{[x_j, \mathcal{A}_\infty[\mathcal{D}](x_j)], \; x_j := \|\mathcal{D}_j\|\}_{j=1}^i \cup \{[\infty, \beta]\} \tag{36}$$

and the observations of the first set are lower or equal than $100 \leq \beta$, we have that the asymptotic backbone $\{\hat{\alpha}_i^\beta\}_{i>\omega}$ of $\hat{\mathcal{A}}^{\frac{\beta}{\pi}}[\mathcal{D}_\sigma^\mathcal{K}]$ is never greater than β and therefore the residuals $\{\hat{\rho}_i^\beta(\infty)\}_{i>\omega}$ are invariably positive or null, because

$$\hat{\rho}_i^\beta(\infty) := \hat{\mathcal{A}}_i^\beta(\infty) - \hat{\alpha}_i^\beta := \beta - \hat{\alpha}_i^\beta \tag{37}$$

As $\{[x_i, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_i)], \; x_i := \|\mathcal{D}_i\|\}_{i \in \mathbb{N}}$ is increasing, its impact on the generation of the learning trends $\{\hat{\mathcal{A}}_i^{\beta}[\mathcal{D}_\sigma^{\mathcal{K}}]\}_{i>\omega}$ grows with each sample and therefore the absolute value of the residuals $\{\hat{\rho}_i^{\beta}(\infty)\}_{i>\omega}$ also. We then derive that:

$$\hat{\mathcal{A}}_i^{\beta}(\infty) - \hat{\mathcal{A}}_{i+1}^{\beta}(\infty) = 0 \geq \hat{\rho}_i^{\beta}(\infty) - \hat{\rho}_{i+1}^{\beta}(\infty), \; \forall i > \omega \tag{38}$$

and Equation (11) verifies.

A.4. Proof of Theorem 7

By Theorem 6, as $\mathcal{A}^{\pi}[\mathcal{D}_\sigma^{\mathcal{K}}]$ is a learning trace with fixed anchor, its asymptotic backbone $\{\alpha_i\}_{i>\omega}$ is decreasing and the thesis is proved applying the same reasoning used in Theorem 3.

A.5. Proof of Theorem 8

Following Theorem 4, as $\{\hat{\mathcal{A}}_i^{\beta,\ell}(\infty)\}_{i>\omega}$ converges to $\hat{\mathcal{A}}_{\omega+\ell}^{\beta,\ell}(\infty)$, it is enough to prove that its hypotheses verify. With regard to the condition expressed in Equation (10), let us assume $\{\hat{\mathcal{A}}_i^{\beta}[\mathcal{D}_\sigma^{\mathcal{K}}]\}_{i>\omega}$ the fixed anchoring learning trace of value β for $\mathcal{A}^{\pi}[\mathcal{D}_\sigma^{\mathcal{K}}]$. As its asymptotic backbone $\{\hat{\alpha}_i^{\beta}\}_{i>\omega}$ is decreasing and $\wp \geq \omega$, we have that, $\forall i \geq \wp + \ell + 1$

$$\hat{\alpha}_{\wp+\ell}^{\beta} \geq \hat{\alpha}_i^{\beta} := \lim_{x \to \infty} \hat{\mathcal{A}}_i^{\beta}[\mathcal{D}_\sigma^{\mathcal{K}}](x) \tag{39}$$

where $\hat{\mathcal{A}}_i^{\beta}[\mathcal{D}_\sigma^{\mathcal{K}}]$ is a curve fitting the set of values

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \hat{\mathcal{A}}_i^{\beta}(\infty)]\} = \{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \beta]\}, \; x_j := \|\mathcal{D}_j\| \tag{40}$$

Since $\mathcal{A}_i^{\pi}[\mathcal{D}_\sigma^{\mathcal{K}}]$ is a curve fitting the first set $\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)], \; x_j := \|\mathcal{D}_j\|\}_{j=1}^{i}$ and by hypothesis $\beta \geq 100 \geq \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j), \forall j \in \mathbb{N}$, we deduce that $\hat{\alpha}_i^{\beta} \geq \alpha_i$, with $\{\alpha_i\}_{i \in \mathbb{N}}$ the asymptotic backbone for the reference $[\mathcal{A}^{\pi}[\mathcal{D}_\sigma^{\mathcal{K}}], \omega]$. Taking also into account that $\lim_{x \to \infty} \hat{\mathcal{A}}_i^{\beta}[\mathcal{D}_\sigma^{\mathcal{K}}](x) := \hat{\alpha}_i^{\beta}$ and $\lim_{x \to \infty} \mathcal{A}_i^{\pi}[\mathcal{D}_\sigma^{\mathcal{K}}](x) := \alpha_i$, it verifies that

$$\hat{\mathcal{A}}_i^{\beta,\ell}(\infty) := \hat{\alpha}_{\wp+\ell}^{\beta,\ell} := \hat{\alpha}_{\wp+\ell}^{\beta} \geq \hat{\alpha}_i^{\beta} \geq \alpha_i, \; \forall i \geq \wp + \ell + 1 \tag{41}$$

Furthermore, since $\wp$ is precisely the first level after the working one ω from which the asymptotic backbone is below 100, we deduce that

$$\hat{\mathcal{A}}_i^{\beta,\ell}(\infty) := \beta \geq 100 \geq \alpha_i, \; \forall \, \wp + \ell + 1 > i > \omega \tag{42}$$

and therefore $\hat{\mathcal{A}}_i^{\beta,\ell}(\infty) \geq \alpha_i, \forall i > \omega$, thus matching the relationship described in Equation (10).

With respect to the condition in Equation (11), we study it separately in each interval of definition for the asymptotic backbone $\{\hat{\alpha}_i^{\beta,\ell}\}_{i>\omega}$. Let us first assume $\wp + \ell + 1 > i > \omega$. As in this case

$$\hat{\mathcal{A}}_i^{\beta,\ell}(\infty) := \hat{\mathcal{A}}_i^{\beta}(\infty) := \beta \geq 100 \tag{43}$$

we can apply the same reasoning used in Theorem 6 to prove that $\{\hat{\alpha}_i^{\beta,\ell}\}_{i>\omega}$ is decreasing in that interval.

Let us now consider $i > \wp + \ell + 1$. Then, $\hat{\mathcal{A}}_i^{\beta,\ell}[\mathcal{D}_\sigma^{\mathcal{K}}]$ is here a curve fitting the collection of values

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \hat{\mathcal{A}}_i^{\beta,\ell}(\infty)]\} = \{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \hat{\alpha}_{\wp+\ell}^{\beta,\ell}]\} \tag{44}$$

with $x_j := \|\mathcal{D}_j\|$. Given that the observations of the first set are always lower or equal than

$$\hat{\mathcal{A}}_i^{\beta,\ell}(\infty) := \hat{\alpha}_{\wp+\ell}^{\beta,\ell} := \hat{\alpha}_{\wp+\ell}^{\beta} \tag{45}$$

because Theorem 6 states that $\{\hat{\alpha}_i^{\beta}\}_{i>\omega}$ is decreasing, the subsequence $\{\hat{\alpha}_i^{\beta,\ell}\}_{i>\wp+\ell+1}$ of the asymptotic backbone of $\hat{\mathcal{A}}_i^{\beta,\ell}[\mathcal{D}_\sigma^{\mathcal{K}}]$ is never greater than $\hat{\alpha}_{\wp+\ell}^{\beta,\ell}$ and therefore the sequence of its associated residuals $\{\hat{\rho}_i^{\beta,\ell}(\infty)\}_{i>\wp+\ell+1}$ is invariably positive or null, because

$$\hat{\rho}_i^{\beta,\ell}(\infty) := \hat{\mathcal{A}}_i^{\beta,\ell}(\infty) - \hat{\alpha}_i^{\beta,\ell} := \hat{\mathcal{A}}_i^{\beta,\ell}(\infty) - \lim_{x \to \infty} \hat{\mathcal{A}}_i^{\beta,\ell}[\mathcal{D}_\sigma^{\mathcal{K}}](x) \tag{46}$$

As $\{[x_i, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_i)], \; x_i := \|\mathcal{D}_i\|\}_{i \in \mathbb{N}}$ is increasing, its impact on the generation of the learning trends $\{\hat{\mathcal{A}}_i^{\beta,\ell}[\mathcal{D}_\sigma^{\mathcal{K}}]\}_{i>\wp+\ell+1}$ grows with each sample and therefore the absolute value of the residuals $\{\hat{\rho}_i^{\beta,\ell}(\infty)\}_{i>\wp+\ell+1}$ also. We then derive that:

$$\hat{\mathcal{A}}_i^{\beta,\ell}(\infty) - \hat{\mathcal{A}}_{i+1}^{\beta,\ell}(\infty) = 0 \geq \hat{\rho}_i^{\beta,\ell}(\infty) - \hat{\rho}_{i+1}^{\beta,\ell}(\infty), \ \forall i > \wp + \ell + 1 \tag{47}$$

and Equation (11) also verifies in the latter interval.

That way, what is lacking to match the relationship in Equation (11) on its full application domain is to demonstrate that

$$\hat{\mathcal{A}}_{\wp+\ell}^{\beta,\ell}(\infty) - \hat{\mathcal{A}}_{\wp+\ell+1}^{\beta,\ell}(\infty) \geq \hat{\rho}_{\wp+\ell}^{\beta,\ell}(\infty) - \hat{\rho}_{\wp+\ell+1}^{\beta,\ell}(\infty) \tag{48}$$

which is equivalent to prove that $\hat{\alpha}_{\wp+\ell}^{\beta,\ell} \geq \hat{\alpha}_{\wp+\ell+1}^{\beta,\ell}$, because $\hat{\rho}_i^{\beta,\ell}(\infty) := \hat{\mathcal{A}}_i^{\beta,\ell}(\infty) - \hat{\alpha}_i^{\beta,\ell}$.

From Theorem 6, $\{\hat{\alpha}_i^{\beta}\}_{i>\omega}$ converges decreasingly to the final accuracy $\alpha_{\boldsymbol{\infty}} := \lim_{x\to\infty} \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x)$ attained from the training process. As $\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}]$ is increasing, we infer that

$$\mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_i) \leq \alpha_{\boldsymbol{\infty}} \leq \hat{\alpha}_{\wp+\ell}^{\beta,\ell} := \hat{\alpha}_{\wp+\ell}^{\beta,\ell}, \ x_i := \|\mathcal{D}_i\| \tag{49}$$

Furthermore, $\hat{\alpha}_{\wp+\ell+1}^{\beta,\ell} := \lim_{x\to\infty} \hat{\mathcal{A}}_{\wp+\ell+1}^{\beta,\ell}[\mathcal{D}_\sigma^{\mathcal{K}}](x)$, where $\hat{\mathcal{A}}_{\wp+\ell+1}^{\beta,\ell}[\mathcal{D}_\sigma^{\mathcal{K}}]$ is a fitting curve for the set

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{\wp+\ell} \cup \{[\infty, \hat{\mathcal{A}}_{\wp+\ell}^{\beta,\ell}(\infty)]\} = \{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{\wp+\ell} \cup \{[\infty, \hat{\alpha}_{\wp+\ell-1}^{\beta,\ell}]\} \tag{50}$$

with $x_j := \|\mathcal{D}_j\|$. As we have established that the ordinates of these values are below $\hat{\alpha}_{\wp+\ell+1}^{\beta,\ell}$, we derive what we were looking for

$$\hat{\alpha}_{\wp+\ell}^{\beta,\ell} \geq \lim_{x\to\infty} \hat{\mathcal{A}}_{\wp+\ell+1}^{\beta,\ell}[\mathcal{D}_\sigma^{\mathcal{K}}](x) := \hat{\alpha}_{\wp+\ell+1}^{\beta,\ell} \tag{51}$$

thus terminating the proof.

A.6. Proof of Theorem 9

We first address the Equations (15) and (16), when no hypothesis regarding the monotony of the asymptotic backbone $\{\alpha_i\}_{i\in\mathbb{N}}$ associated to the reference is established. Given that the former has already been stated [42], we focus on the second one, starting from its left-most inequality.

Let us assume $\jmath > \iota$, $\eta > \beta$ and $i \geq \wp + \jmath$, then $\hat{\alpha}_i^{\beta,\iota} := \lim_{x\to\infty} \hat{\mathcal{A}}_i^{\beta,\iota}[\mathcal{D}_\sigma^{\mathcal{K}}](x)$, with $\hat{\mathcal{A}}_i^{\beta,\iota}[\mathcal{D}_\sigma^{\mathcal{K}}]$ a fitting curve for the values

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}_\sigma^{\mathcal{K}}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \hat{\mathcal{A}}_i^{\beta,\iota}(\infty)]\} = \{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}_\sigma^{\mathcal{K}}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \hat{\alpha}_i^{\beta}]\} \tag{52}$$

with $x_j := \|\mathcal{D}_j\|$. Similarly, $\hat{\alpha}_i^{\beta,\jmath} := \lim_{x\to\infty} \hat{\mathcal{A}}_i^{\beta,\jmath}[\mathcal{D}_\sigma^{\mathcal{K}}](x)$, with $\hat{\mathcal{A}}_i^{\beta,\jmath}[\mathcal{D}_\sigma^{\mathcal{K}}]$ a curve fitting

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)], \ x_j := \|\mathcal{D}_j\|\}_{j=1}^{i} \cup \{[\infty, \hat{\alpha}_j^{\beta}]\} \tag{53}$$

where, as by Theorem 6 the sequence $\{\hat{\alpha}_i^{\beta}\}_{i>\omega}$ converges decreasingly to $\alpha_\infty = \alpha_{\boldsymbol{\infty}} \geq 0$, we conclude that $\hat{\alpha}_i^{\beta} > \hat{\alpha}_j^{\beta}$. As Theorem 8 guarantees that $\{\hat{\alpha}_i^{\beta,\iota}\}_{i>\omega}$ and $\{\hat{\alpha}_i^{\beta,\jmath}\}_{i>\omega}$ also converge decreasingly to $\alpha_\infty = \alpha_{\boldsymbol{\infty}} \geq 0$, and $\wp \geq \omega$, we can conclude that $|\hat{\alpha}_i^{\beta,\jmath} - \alpha_{\boldsymbol{\infty}}| \leq |\hat{\alpha}_i^{\beta,\iota} - \alpha_{\boldsymbol{\infty}}|$. We thereby demonstrate the inequality in question.

On the other hand, $\hat{\alpha}_i^{\eta,\iota} := \lim_{x\to\infty} \hat{\mathcal{A}}_i^{\eta,\iota}[\mathcal{D}_\sigma^{\mathcal{K}}](x)$, with $\hat{\mathcal{A}}_i^{\eta,\iota}[\mathcal{D}_\sigma^{\mathcal{K}}]$ a fitting curve for the values

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \hat{\mathcal{A}}_i^{\eta,\iota}(\infty)]\} = \{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \hat{\alpha}_i^{\eta}]\} \tag{54}$$

where $x_j := \|\mathcal{D}_j\|$ and $\hat{\alpha}_i^{\eta} := \lim_{x\to\infty} \hat{\mathcal{A}}_i^{\eta}[\mathcal{D}_\sigma^{\mathcal{K}}](x)$, with $\hat{\mathcal{A}}_i^{\eta}[\mathcal{D}_\sigma^{\mathcal{K}}]$ a fitting curve for the values

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \hat{\mathcal{A}}_i^{\eta}(\infty)]\} = \{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}_{j=1}^{i} \cup \{[\infty, \eta]\}, \ x_j := \|\mathcal{D}_j\| \tag{55}$$

Similarly, $\hat{\alpha}_i^{\beta} := \lim_{x\to\infty} \hat{\mathcal{A}}_i^{\beta}[\mathcal{D}_\sigma^{\mathcal{K}}](x)$, with $\hat{\mathcal{A}}_i^{\beta}[\mathcal{D}_\sigma^{\mathcal{K}}]$ a fitting curve for the values

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)], \ x_j := \|\mathcal{D}_j\|\}_{j=1}^{i} \cup \{[\infty, \beta]\} \tag{56}$$

Since $\eta > \beta$, we then have that $\hat{\alpha}_i^{\eta} \leq \hat{\alpha}_i^{\beta}$, and therefore $\hat{\alpha}_i^{\eta,\iota} \leq \hat{\alpha}_i^{\beta,\iota}$. Accordingly, we also conclude that $|\hat{\alpha}_i^{\beta,\iota} - \alpha_{\boldsymbol{\infty}}| \leq |\hat{\alpha}_i^{\eta,\iota} - \alpha_{\boldsymbol{\infty}}|$, because $\{\hat{\alpha}_i^{\eta,\iota}\}_{i>\omega}$ and $\{\hat{\alpha}_i^{\beta,\iota}\}_{i>\omega}$ are both positive definite and $\wp \geq \omega$.

Finally, we demonstrate the Equation (17), when the asymptotic backbone $\{\alpha_i\}_{i\in\mathbb{N}}$ of the reference is decreasing. On the basis of the result previously stated for the generic case, and taking into account that by definition $\omega \leq \wp \leq 100 \leq \beta$ and $\hat{\mathcal{A}}^{\beta,0}_{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}] = \hat{\mathcal{A}}^{\beta}_{\pi}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$, it is sufficient to establish that

$$| \hat{\alpha}_i - \alpha_{\boldsymbol{\infty}} | \leq | \hat{\alpha}^{\beta}_i - \alpha_{\boldsymbol{\infty}} |, \ \forall i > \wp \tag{57}$$

where $\hat{\alpha}_i := \lim_{x\to\infty} \hat{\mathcal{A}}^{\pi}_i[\mathcal{D}^{\mathcal{K}}_{\sigma}](x)$, with $\hat{\mathcal{A}}^{\pi}_i[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ a fitting curve for the values

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}^i_{j=1} \ \cup \ \{[\infty, \hat{\mathcal{A}}_i(\infty)]\} = \{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}^i_{j=1} \ \cup \ \{[\infty, \hat{\alpha}_{i-1}]\} \tag{58}$$

with $x_j := \|\mathcal{D}_j\|$. Furthermore, $\hat{\alpha}^{\beta}_i := \lim_{x\to\infty} \hat{\mathcal{A}}^{\beta}_{\overline{\pi}}[\mathcal{D}^{\mathcal{K}}_{\sigma}](x)$, with $\hat{\mathcal{A}}^{\beta}_{\overline{\pi}}[\mathcal{D}^{\mathcal{K}}_{\sigma}]$ a curve fitting

$$\{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}^i_{j=1} \ \cup \ \{[\infty, \hat{\mathcal{A}}^{\beta}_i(\infty)]\} = \{[x_j, \mathcal{A}_{\boldsymbol{\infty}}[\mathcal{D}](x_j)]\}^i_{j=1} \ \cup \ \{[\infty, \beta]\}, \ x_j := \|\mathcal{D}_j\| \tag{59}$$

So, as $\{\hat{\alpha}_i\}_{i>\omega}$ is positive definite, for stating the desired relation it is enough to prove that $\hat{\alpha}_{i-1} \leq \beta$, which is trivial because $\hat{\alpha}_{i-1} \leq 100 \leq \beta$. This completes the proof.

References

[1] T.M. Apostol, Mathematical Analysis, Narosa Book Distributors Pvt Ltd, New Delhi, 2000.
[2] E. Bauer, R. Kohavi, An empirical comparison of voting classification algorithms: bagging, boosting, and variants, Mach. Learn. 36 (1–2) (Jul. 1999) 105–139.
[3] M. Bloodgood, K. Vijay-Shanker, A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping, in: Proceedings of the 13th Conference on Computational Natural Language Learning, Boulder, 2009, pp. 39–47.
[4] M.A. Branch, T.F. Coleman, Y. Li, A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems, SIAM J. Sci. Comput. 21 (1) (1999) 1–23.
[5] T. Brants, TnT: a statistical part-of-speech tagger, in: Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, 2000, pp. 224–231.
[6] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[7] E. Brill, Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging, Comput. Linguist. 21 (4) (1995) 543–565.
[8] G. Chrupala, G. Dinu, J. van Genabith, Learning morphology with Morfette, in: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, 2008, pp. 2362–2367.
[9] A. Clark, C. Fox, S. Lappin, The Handbook of Computational Linguistics and Natural Language Processing, John Wiley & Sons, Hoboken, 2010.
[10] M. Collins, Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (Vol. 10), Philadelphia, 2002, pp. 1–8.
[11] W. Daelemans, J. Zavrel, P. Berck, S. Gillis, MBT: a memory–based part-of-speech tagger generator, in: Proceedings of the 4th Workshop on Very Large Corpora, Copenhagen, 1996, pp. 14–27.
[12] C. Domingo, R. Gavaldà, O. Watanabe, Adaptive sampling methods for scaling up knowledge discovery algorithms, Data Min. Knowl. Discov. 6 (2) (2002) 131–152.
[13] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proceedings of the 13th International Conference on Machine Learning, Bari, 1996, pp. 148–156.
[14] L. Frey, D. Fischer, Modeling decision tree performance with the power law, in: Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, 1999, pp. 59–65.
[15] N. García-Pedrajas, A. De Haro-García, Boosting instance selection algorithms, Knowl.-Based Syst. 67 (Sep. 2014) 342–360.
[16] E. Giesbrecht, S. Evert, Is part-of-speech tagging a solved task? An evaluation of POS taggers for the German web as corpus, in: Proceedings of the 5th Web as Corpus Workshop, San Sebastian, 2009, pp. 27–35.
[17] J. Giménez, L. Márquez, SVMTool: a general POS tagger generator based on support vector machines, in: Proceedings of the 4th International Conference on Language Resources and Evaluation, Lisbon, 2004, pp. 43–46.
[18] L. Hinrichs, N. Smith, B. Waibel, Manual of information for the part-of-speech-tagged, post-edited 'Brown' corpora, ICAME J. 34 (2010) 189–233.
[19] R.A. Howard, Decision analysis: applied decision theory, in: Proceedings of the 4th International Conference on Operational Research, Cambridge, 1966, pp. 55–71.
[20] G. John, P. Langley, Static versus dynamic sampling for data mining, in: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, 1996, pp. 367–370.
[21] A. Kapoor, R. Greiner, Learning and classifying under hard budgets, in: Machine Learning: ECML 2005, Springer-Verlag, 2005, pp. 170–181.
[22] M. Last, Improving data mining utility with projective sampling, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, 2009, pp. 487–496.
[23] R. Leite, P. Brazdil, An iterative process for building learning curves and predicting relative performance of classifiers, in: Proceedings of the Artificial Intelligence 13th Portuguese Conference on Progress in Artificial Intelligence, Guimarães, 2007, pp. 87–98.
[24] K.T. Leung, D.S. Parker, Empirical comparisons of various voting methods in bagging, in: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, 2003, pp. 595–600.
[25] D.D. Lewis, W.A. Gale, A sequential algorithm for training text classifiers, in: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994, pp. 3–12.
[26] C. Mair, G. Leech, The Freiburg-Brown corpus ('Frown') (POS-tagged version). Albert-Ludwigs-Universität Freiburg and University of Lancaster, Available for academic use through http://clu.uni.no/icame/, 2007.
[27] M.P. Marcus, B. Santorini, M.A. Marcinkiewicz, A. Taylor, Treebank-3 LDC99T42. Web Download File, Linguistic Data Consortium, Philadelphia, 1999.
[28] C. Meek, B. Thiesson, D. Heckerman, The learning-curve sampling method applied to model-based clustering, J. Mach. Learn. Res. 2 (Mar. 2002) 397–418.
[29] G. Ngai, R. Florian, Transformation-based learning in the fast lane, in: Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, Pittsburgh, 2001, pp. 1–8.

[30] F. Provost, D. Jensen, T. Oates, Efficient progressive sampling, in: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, 1999, pp. 23–32.

[31] A. Ratnaparkhi, A maximum entropy model for part-of-speech tagging, in: Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing, Philadelphia, 1996, pp. 133–142.

[32] R. Reichart, O. Abend, A. Rappoport, Type level clustering evaluation: new measures and a pos induction case study, in: Proceedings of the Fourteenth Conference on Computational Natural Language Learning, CoNLL '10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 77–87.

[33] R.E. Schapire, The strength of weak learnability, Mach. Learn. 5 (2) (1990) 197–227.

[34] H. Schmid, Probabilistic part-of-speech tagging using decision trees, in: Proceedings of the International Conference on New Methods in Language Processing, Manchester, 1994, pp. 44–49.

[35] H. Schmid, F. Laws, Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging, in: Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1, COLING '08, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 777–784.

[36] H. Schütze, E. Velipasaoglu, J.O. Pedersen, Performance thresholding in practical text classification, in: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Arlington, 2006, pp. 662–671.

[37] H.-J. Song, J.-W. Son, T.-G. Noh, S.-B. Park, S.-J. Lee, A cost sensitive part-of-speech tagging: differentiating serious errors from minor errors, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers (Vol. 1), Jeju, Island, 2012, pp. 1025–1034.

[38] K. Tomanek, U. Hahn, Approximating learning curves for active-learning-driven annotation, in: Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, 2008, pp. 1319–1324.

[39] K. Toutanova, D. Klein, C.D. Manning, Y. Singer, Feature-rich part-of-speech tagging with a cyclic dependency network, in: Proceedings of the 2003 Annual Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (Vol. 1), Edmonton, 2003, pp. 173–180.

[40] Y. Tsuruoka, Y. Miyao, J. Kazama, Learning with lookahead: can history-based models rival globally optimized models?, in: Proceedings of the 15th Conference on Computational Natural Language Learning, Portland, 2011, pp. 238–246.

[41] H. van Halteren, Performance of taggers, in: Syntactic Wordclass Tagging, Kluwer Academic Pub., Hingham, 1999, pp. 81–94.

[42] M. Vilares, V.M. Darriba, F.J. Ribadas, Modeling of learning curves with applications to pos tagging, Comput. Speech Lang. 41 (2017) 1–28.

[43] A. Vlachos, A stopping criterion for active learning, Comput. Speech Lang. 22 (3) (2008) 295–312.

[44] G. Weiss, Y. Tian, Maximizing classifier utility when there are data acquisition and modeling costs, Data Min. Knowl. Discov. 17 (2) (2008) 253–282.