Adaptive Sampling Under Low Noise Conditions

Nicolò Cesa-Bianchi

Adaptive Sampling Under Low Noise Conditions[1]

Nicolò Cesa-Bianchi

Dipartimento di Scienze dell'Informazione
Università degli Studi di Milano
20135 Milano, Italy

Abstract: We survey some recent results on efficient margin-based algorithms for adaptive sampling in binary classification tasks. Using the so-called Mammen-Tsybakov low noise condition to parametrize the distribution of covariates, and assuming linear label noise, we state bounds on the convergence rate of the adaptive sampler to the Bayes risk. These bounds show that, excluding logarithmic factors, the average risk converges to the Bayes risk at rate $N^{-(1+\alpha)(2+\alpha)/2(3+\alpha)}$, where N denotes the number of queried labels and α is the nonnegative exponent in the low noise condition. For all $\alpha > \sqrt{3} - 1$ this convergence rate is asymptotically faster than the rate $N^{-(1+\alpha)/(2+\alpha)}$ achieved by the fully supervised version of the base adaptive sampler, which queries all labels. Moreover, for $\alpha \to \infty$ (hard margin condition) the gap between the semi- and fully-supervised rates becomes exponential.

Keywords: linear classification, regularized least squares, regret, convergence rates

In the online learning model for binary classification the learner receives a sequence of instances generated by an unknown source. Each time a new instance is received the learner predicts its binary label, which is then immediately disclosed before the next instance is observed. This protocol is natural in many applications, for instance weather forecasting or stock market prediction, because Nature (or the market) is spontaneously revealing the true label after each learner's guess. However, in many other cases obtaining labels may be an expensive process.

In order to address this problem selective sampling has been proposed as a more realistic alternative. In selective sampling the true label of the current instance is never revealed unless the learner decides to issue an explicit query. The learner's performance is then measured with respect to both the number of mistakes (made on the entire sequence of instances) and the number of queries.

A natural goal in this setting is trying to sample labels at a rate guaranteeing a certain level of predictive accuracy. To achieve this, a strategy may combine a measure of utility of examples with a measure of confidence for the current prediction. In the case of learning

[1]This extended abstract is based on work, co-authored with G. Cavallanti and C. Gentile, submitted for journal publication (G. Cavallanti, N. Cesa-Bianchi, and C. Gentile, Learning Noisy Linear Classifiers via Selective Sampling, 2009).

The author gratefully acknowledges partial support by the PASCAL2 Network of Excellence under EC grant no. 216886. This publication only reflects the author's views.

with linear functions, a statistic that has often been used to quantify both utility and confidence is the margin.

In this abstract we describe our recent work [5] in which the selective sampling rule queries a label whenever the margin of the current instance, with respect to the current linear hypothesis, is smaller (in absolute value) than a suitable adaptive threshold. Technical reasons prevent us from analyzing the sampler that exactly queries the label of the instance whose small margin was observed, which is intuitively the right thing to do in order to maximize the information conveyed by the label. Instead, we consider a sampler that queries the label of the next instance generated by the random process, irrespective to its actual margin. By doing this, we are able to measure the advantage provided by sampling at the rate at which we observe small margins. However, we can not say anything about the advantage, confirmed by the experiments in [5], of querying precisely the small margin instances. In order to emphasize that we do not select which instances to sample, but rather adapt our sampling rate to the distribution of observed margins, we call our approach *adaptive sampling*.

We consider data $(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots$ with binary response variables $Y_t \in \{-1, +1\}$ (*labels*) and d-dimensional covariates $\boldsymbol{X}_t \in \mathbb{R}^d$ (*instances*). We assume instances $\boldsymbol{X}_t$ are drawn i.i.d. from an unknown distribution on the surface of the unit Euclidean ball in $\mathbb{R}^d$, so that $\|\boldsymbol{X}_t\| = 1$ w.p. 1 for all $t \geq 1$. We also assume that the conditional distribution of labels Y_t satisfies $\mathbb{E}\big[Y_t \,\big|\, \boldsymbol{X}_t = \boldsymbol{x}_t\big] = \boldsymbol{u}^\top \boldsymbol{x}_t$ for all $t \geq 1$, where $\boldsymbol{u} \in \mathbb{R}^d$ is a fixed and unknown unit-norm vector. Hence, $\textsc{sgn}(f^*)$, for $f^*(\boldsymbol{X}) = \boldsymbol{u}^\top \boldsymbol{X}$, is the Bayes optimal classifier for the above data model. In what follows, all probabilities $\mathbb{P}$ and expectations $\mathbb{E}$ are understood with respect to the joint distribution of the data. Let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary measurable function. The *instantaneous regret* $R(f)$ is the excess risk of $\textsc{sgn}(f)$ with respect to the Bayes risk, that is, $R(f) = \mathbb{P}(Y_1\, f(\boldsymbol{X}_1) < 0) - \mathbb{P}(Y_1\, f^*(\boldsymbol{X}_1) < 0)$. Let $f_1, f_2, \ldots$ be a sequence of real functions where each f_t is measurable with respect to the σ-algebra $\mathcal{F}_{t-1}$ generated by $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{t-1}, Y_{t-1}), \boldsymbol{X}_t$. Let $R_{t-1}(f_t)$ be the *conditional* instantaneous regret $R_{t-1}(f_t) = \mathbb{P}_{t-1}(Y_1\, f_t(\boldsymbol{X}_t) < 0) - \mathbb{P}_{t-1}(Y_1\, f^*(\boldsymbol{X}_t) < 0)$ where $\mathbb{P}_t$ denotes $\mathbb{P}$ conditioned on $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_t, Y_t)$. The goal is to bound the *cumulative (expected) regret*

$$\mathbb{E}\left[\sum_{t=1}^{n} R(f_t)\right] = \mathbb{E}\big[R_0(f_1) + R_1(f_2) + \cdots + R_{n-1}(f_n)\big]$$

as a function of n, and other relevant quantities. Observe that, although the learner's predictions can only depend on the observed instances and queried labels, the above regret is computed over *all* time steps, including those time steps t when the adaptive sampler did not issue a query.

We consider algorithms that predict the value of Y_t through $\textsc{sgn}\big(\widehat{\Delta}_t\big)$, where the margin $\widehat{\Delta}_t = \boldsymbol{w}_t^\top \boldsymbol{x}_t$ at time t is based on the regularized least squares (RLS) estimator $\boldsymbol{W}_t$ defined over the set of previously queried examples. More precisely, let N_{t-1} be the

number of queried examples in the first $t-1$ steps, let $S_{t-1} = \big[\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_{N_{t-1}}\big]$ be the $d \times N_{t-1}$ matrix of their instances, and let $\boldsymbol{y}_{t-1} = \big(y'_1, \ldots, y'_{N_{t-1}}\big)$ be the vector of the corresponding labels. Then

$$\boldsymbol{W}_t = \big(I + S_{t-1}\, S_{t-1}^\top + \boldsymbol{X}_t\, \boldsymbol{X}_t^\top\big)^{-1} S_{t-1}\, \boldsymbol{Y}_{t-1} \tag{1}$$

where I is the $d \times d$ identity matrix. Note that $\boldsymbol{W}_t$ defined by (1) depends on the current instance $\boldsymbol{X}_t$. The regularized least squares estimator in this particular form has been considered by Vovk [10] and, independently, by Azoury and Warmuth [1].

We denote by Δ_t the Bayes margin $f^*(\boldsymbol{X}_t) = \boldsymbol{u}^\top \boldsymbol{X}_t$. Note that $\widehat{\Delta}_t$ is measurable with respect to the σ-algebra generated by $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{t-1}, Y_{t-1}), \boldsymbol{X}_t$.

We model the distribution of the instances around the hyperplane $\boldsymbol{u}^\top \boldsymbol{x} = 0$, using the popular *Mammen-Tsybakov low noise condition* [9]:

There exist $c > 0$ and $\alpha \geq 0$ such that $\mathbb{P}\big(|f^*(\boldsymbol{X}_1)| < \varepsilon\big) \leq c\, \varepsilon^\alpha$ for all $\varepsilon > 0$.

It can be shown that $\alpha \to \infty$ implies the *hard margin condition*. Namely, $\big|f^*(\boldsymbol{X}_1)\big| \geq 1/(2c)$ with probability 1.

We now state a regret bound for the fully supervised version of our sampling algorithm This algorithm predicts using the RLS estimate (1) and queries the label of every observed instance.

Theorem 1 *Assume the Mammen-Tsybakov low noise condition holds with exponent $\alpha \geq 0$ and constant $c > 0$. Then the expected cumulative regret after n steps of the fully supervised algorithm is bounded by*

$$\mathbb{E}\left[\Big(4c\big(1 + \ln\big|I + S_n S_n^\top\big|\big)\Big)^{\frac{1+\alpha}{2+\alpha}}\right] n^{\frac{1}{2+\alpha}} \ .$$

This, in turn, is upper bounded by

$$\left[4c\left(1 + \sum_{i=1}^{d} \ln(1 + n\lambda_i)\right)\right]^{\frac{1+\alpha}{2+\alpha}} n^{\frac{1}{2+\alpha}} = O\left(\big(d \ln n\big)^{\frac{1+\alpha}{2+\alpha}} n^{\frac{1}{2+\alpha}}\right) \ .$$

In the above $|\cdot|$ denotes the determinant, $S_n = \big[\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\big]$ is the (random) matrix containing all instances, and λ_i is the ith eigenvalue of $\mathbb{E}\big[\boldsymbol{X}_1 \boldsymbol{X}_1^\top\big]$.

When $\alpha = 0$ (corresponding to a vacuous noise condition) the bound of Theorem 1 reduces to $O\big(\sqrt{d\, n \ln n}\big)$. When $\alpha \to \infty$ (the hard margin condition) the bound gives the logarithmic behavior $O\big(d \ln n\big)$. Note that $\sum_{i=1}^{d} \ln(1 + n\lambda_i)$ is substantially smaller than $d \ln n$ whenever the spectrum of $\mathbb{E}[\boldsymbol{X}_1 \boldsymbol{X}_1^\top]$ is rapidly decreasing. In fact, the second bound is clearly meaningful even when $d = \infty$, while the third one only applies to the finite dimensional case.

We now turn to the description of our adaptive sampling algorithm based on the same RLS estimate (1). The sampler issues a query at time t based on both the query counter N_{t-1} and the (signed) margin $\widehat{\Delta}_t$. Specifically, if $N_{t-1} \leq (128 \ln t)/\big(\lambda \widehat{\Delta}_t^2\big)$ or $N_{t-1} \leq \big(16\lambda^2\big) \max\{d, \ln t\}$, then the label of the next instance $\boldsymbol{X}_{t+1}$ is queried. Here λ denotes the smallest eigenvalue of the process correlation matrix $\mathbb{E}\big[\boldsymbol{X}_1 \boldsymbol{X}_1^\top\big]$ (this eigenvalue must thus be known to the algorithm). The first condition detects whether the number N_{t-1} of queried labels is smaller than our current estimate of $1/\Delta_t^2$ at confidence level $1-1/t$. The second condition ensures that the smallest eigenvalue of the empirical correlation matrix $S_t\, S_t^\top / N_t$ converges to the smallest eigenvalue of the process correlation matrix. This is needed to control bias and variance of the margin estimate $\widehat{\Delta}_t$ and requires $\mathbb{E}\big[\boldsymbol{X}_1 \boldsymbol{X}_1^\top\big]$ to be full rank (this requirement can be dropped at the cost of a more involved analysis —see discussion in [5]). Finally, observe that once an example is scheduled to be queried the algorithm cannot change its mind on the next time step (because, say, the new margin now happens to be larger than the current threshold).

The following result shows that our algorithm is able to adaptively optimize the sampling rate by exploiting the additional information provided by the examples having small margin. The appropriate rate clearly depends on the (unknown) amount of noise α in the Mammen-Tsybakov condition which the algorithm implicitly learns.

Theorem 2 *Assume the Mammen-Tsybakov low noise condition holds with unknown exponent $\alpha \geq 0$. Then the expected cumulative regret after of the adaptive sampler after n steps is bounded by*

$$O\left(\frac{d + \ln n}{\lambda^2} + \Big(\frac{\ln n}{\lambda}\Big)^{\frac{1+\alpha}{3+\alpha}} n^{\frac{2}{3+\alpha}}\right)$$

whereas the expected number of queried labels is bounded by

$$O\left(\frac{d + \ln n}{\lambda^2} + \Big(\frac{\ln n}{\lambda}\Big)^{\frac{\alpha}{2+\alpha}} n^{\frac{2}{2+\alpha}}\right) .$$

A few remarks. First, as expected, when we compare our semi-supervised adaptive sampler (Theorem 2) to its fully supervised counterpart (Theorem 1), we see that the average instantaneous regret of the former vanishes at a significantly slower rate than the latter, i.e., $n^{-\frac{1+\alpha}{3+\alpha}}$ vs. $n^{-\frac{1+\alpha}{2+\alpha}}$ excluding log factors. Note, however, that the instantaneous regret of the semi-supervised algorithm vanishes faster than the fully-supervised algorithm when both regrets are expressed in terms of the number N of issued queries. To see this consider first the case $\alpha \to \infty$ (the hard margin case). Then both algorithms have an average regret of order $(\ln n)/n$. However, since the semi-supervised algorithm makes only $N = O(\ln n)$ queries, we have that, as a function of N, the average regret of the semi-supervised algorithm is of order N/e^N whereas the fully supervised has only $(\ln N)/N$. We have thus recovered the exponential advantage observed in previous works [4, 6, 7]. When $\alpha = 0$ (vacuous noise conditions), the average regret rates in terms of N become (excluding logarithmic factors) of order $N^{-1/3}$ in the semi-supervised case and of order

$N^{-1/2}$ in the fully supervised case. Hence, there is be a critical value of α where the semi-supervised bound becomes better. In order to find this critical value we write the rates of the average instantaneous regret for $0 \leq \alpha < \infty$ obtaining $N^{-\frac{(1+\alpha)(2+\alpha)}{2(3+\alpha)}}$ (semi-supervised algorithm) and $N^{-\frac{1+\alpha}{2+\alpha}}$ (fully supervised algorithm). By comparing the two exponents we find that, asymptotically, the semi-supervised rate is better than the fully supervised one for all values of $\alpha > \sqrt{3} - 1$. This indicates that adaptive sampling is advantageous when the noise level (as modeled by the Mammen-Tsybakov condition) is not too high.

Second, note that the query rule used by our adaptive sampling algorithm explicitly depends, through λ, on additional information about the data process. This additional information is needed because, unlike the fully supervised classifier of Theorem 1, the adaptive sampler queries labels at random steps. This prevents us from bounding the sum of conditional variances of the RLS estimator through $\ln\left|I + S_n\, S_n^\top\right|$, as we do when proving Theorem 1. Instead, we have to individually bound each conditional variance term via the smallest empirical eigenvalue of the correlation matrix, and this causes the bound of Theorem 2 to depend (inversely) on the smallest process eigenvalue, rather than the whole process eigenspectrum as in Theorem 1.

To our knowledge, the result closest to our work is Ying and Zhou [11], where the authors prove bounds for online linear classification using the same Mammen-Tsybakov low noise condition, though under different distributional assumptions. Note also that fast rates of convergence (i.e., rates faster than $n^{-1/2}$) are typically proven for batch-style algorithms, such as empirical risk minimizers and SVM (see, e.g., [2, 8, 9]; see also [3] for a survey) rather than for online algorithms.

When divided by the number n of steps, the bound of Theorem 1 is of the order $n^{-\frac{1+\alpha}{2+\alpha}}$. Despite the fact we do not have a lower bounding argument holding for our *specific* label noise model $\mathbb{E}\left[Y_t \,\middle|\, \boldsymbol{X}_t\right] = \Delta_t$, we would like to stress that these convergence rates actually match, up to log-factors, the best known upper bounds holding under low noise conditions. Hence, we tend to consider the cumulative rate $n^{\frac{1}{2+\alpha}}$ in Theorem 1 as a good reference result to compare against.

Finally, the second bound in Theorem 1 makes explicit the dependence on the spectrum $\lambda_1, \lambda_2, \ldots$ of the process correlation matrix $\mathbb{E}\left[\boldsymbol{X}_1 \boldsymbol{X}_1^\top\right]$. As far as we can tell, this bound is novel. The analysis of the adaptive sampling algorithm, instead, does not exhibit such a clean dependence on the process spectrum.

References

[1] K.S. Azoury and M.K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[2] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[3] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probability and Statistics*, 9:323–375, 2005.

[4] R. Castro and R.D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

[5] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Learning noisy linear classifiers via selective sampling. Submitted, 2009.

[6] S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of Perceptron-based active learning. In *Proceedings of the 18th Conference on Learning Theory (Colt 2005)*, pages 249–263. Springer, 2005.

[7] Y. Freund, S. Seung, E. Shamir, and N. Tishby. Selective sampling using the Query by Committee algorithm. *Machine Learning*, 28(2/3):133–168, 1997.

[8] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35:575–60, 2007.

[9] A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[10] V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

[11] Y. Ying and D.X. Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52:4775–4788, 2006.