New bounds on the price of bandit feedback for mistake-bounded online multiclass learning

Philip M. Long

Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043 USA. Email: plong@google.com.

Abstract

This paper is about two generalizations of the mistake bound model to online multiclass classification. In the *standard model*, the learner receives the correct classification at the end of each round, and in the *bandit model*, the learner only finds out whether its prediction was correct or not. For a set F of multiclass classifiers, let $\mathrm{opt}_{\mathrm{std}}(F)$ and $\mathrm{opt}_{\mathrm{bandit}}(F)$ be the optimal bounds for learning F according to these two models. We show that an

$$\mathrm{opt}_{\mathrm{bandit}}(F) \leq (1 + o(1))(|Y| \ln |Y|)\mathrm{opt}_{\mathrm{std}}(F)$$

bound is the best possible up to the leading constant, closing a $\Theta(\log |Y|)$ factor gap.

Keywords: online learning, bandit feedback, mistake-bound model, learning theory
68T05

1. Introduction

There are two natural ways to generalize the mistake-bound model [17] to multiclass classification [5].

In the *standard model*, for a set F of functions from some set X to a finite set Y, for an arbitrary $f \in F$ that is unknown to the algorithm, learning proceeds in rounds, and in round t, the algorithm

- receives $x_t \in X$,
- predicts $\hat{y}_t \in Y$, and
- gets $f(x_t)$.

The goal is to bound the number of prediction mistakes in the worst case, over all possible $f \in F$ and $x_1, x_2, ... \in X$.

The *bandit model* [12, 11, 14] (called "weak reinforcement" in [5, 4]) is like the standard model, except that, at the end of each round, the algorithm only finds out whether $\hat{y}_t = f(x_t)$ or not.

Obviously, $\mathrm{opt}_{\mathrm{std}}(F) \leq \mathrm{opt}_{\mathrm{bandit}}(F)$. It is known [4] that, for all F,

$$\mathrm{opt}_{\mathrm{bandit}}(F) \leq (2.01 + o(1)) \, (|Y| \ln |Y|) \, \mathrm{opt}_{\mathrm{std}}(F), \tag{1}$$

and that, for any k and M, there is a set F of functions from a set X to a set Y of size k such that $\mathrm{opt}_{\mathrm{std}}(F) = M$ and

$$\mathrm{opt}_{\mathrm{bandit}}(F) \geq (|Y| - 1)\,\mathrm{opt}_{\mathrm{std}}(F),$$

so that (1) cannot be improved by more than a log factor.

This note shows that, for all $M > 1$ and infinitely many k, there is a set F of functions from a set X to a set Y of size k such that $\mathrm{opt}_{\mathrm{std}}(F) = M$ and

$$\mathrm{opt}_{\mathrm{bandit}}(F) \geq (1 - o(1))\,(|Y| \ln |Y|)\,\mathrm{opt}_{\mathrm{std}}(F), \qquad (2)$$

and that an

$$\mathrm{opt}_{\mathrm{bandit}}(F) \leq (1 + o(1))\,(|Y| \ln |Y|)\,\mathrm{opt}_{\mathrm{std}}(F) \qquad (3)$$

bound holds for all F.

Previous work. In addition to the bounds described above, on-line learning with bandit feedback, side-information and adversarially chosen examples has been heavily studied (see [15, 3, 1, 2, 16, 10, 8, 11]). Daniely and Halbertal [13] studied the price of bandit feedback in the agnostic on-line model, where there is not necessarily an $f \in F$ that always provides the correct classification, and the online learning algorithm is evaluated by comparing its number of mistakes with the best mistake count possible in hindsight obtained by repeatedly applying a classifier in F. The proof of (2) uses analytical tools that were previously used for experimental design [21, 22], and hashing, derandomization and cryptography [9, 19]. The proof of (3) uses tools based on the Weighted Majority algorithm [18, 4].

2. Preliminaries and main results

2.1. Definitions

Define $\mathrm{opt}_{\mathrm{bs}}(k, M)$ to be the best possible bound on $\mathrm{opt}_{\mathrm{bandit}}(F)$ in terms of $M = \mathrm{opt}_{\mathrm{std}}(F)$ and $k = |Y|$. In other words, $\mathrm{opt}_{\mathrm{bs}}(k, M)$ is the maximum, over sets X and sets F of functions from X to $\{0, ..., k - 1\}$ such that $\mathrm{opt}_{\mathrm{std}}(F) = M$, of $\mathrm{opt}_{\mathrm{bandit}}(F)$.

We denote the limit supremum by $\overline{\lim}$.

2.2. Result

The following is our main result.

Theorem 1.

$$\overline{\lim}_{M\to\infty}\overline{\lim}_{k\to\infty}\frac{\mathrm{opt}_{\mathrm{bs}}(k, M)}{kM \ln k} = 1.$$

2.3. The extremal case

For any prime p, let $F_L(p, n)$ be the set of all linear functions from $\{0, ..., p - 1\}^n$ to $\{0, ..., p - 1\}$, where operations are done with respect the finite field $GF(p)$.

In other words, for each $\mathbf{a} \in \{0, ..., p - 1\}^n$, let $f_{\mathbf{a}} : \{0, ..., p - 1\}^n \to \{0, ..., p - 1\}$ be defined by

$$f_{\mathbf{a}}(\mathbf{x}) = (\mathbf{a} \cdot \mathbf{x}) \mod p$$

and let $F_L(p, n) = \{f_{\mathbf{a}} : \mathbf{a} \in \{0, ..., p-1\}^n\}$.

The fact that

$$\mathrm{opt}_{\mathrm{std}}(F_L(p, n)) = n \tag{4}$$

for all primes $p \geq 2$ is essentially known (see [23, 5, 6]). (An algorithm can achieve a mistake bound of n by exploiting the linearity of the target function to always predict correctly whenever $\mathbf{x}_t$ is in the span of previously seen examples. An adversary can force mistakes on any linearly independent set of the domain by answering whichever of 0 or 1 is different from the algorithm's prediction.)

3. Lower bounds

Our lower bound proof will use an adversary that maintains a *version space* [20], a subset of $F_L(p, n)$ that could still be the target. To keep the version space large no matter what the algorithm predicts, the adversary chooses a $\mathbf{x}_t$ for round t that divides it evenly. The first lemma analyzes its ability to do this.

Lemma 1. *For any $S \subseteq \{1, ..., p-1\}^n$, there is a $\mathbf{u}$ such that for all $z \in \{0, ..., p-1\}$,*

$$|\{\mathbf{s} \in S : \mathbf{s} \cdot \mathbf{u} = z \mod p\}| \leq |S|/p + 2\sqrt{|S|}.$$

Lemma 1 is similar to analyses of hashing (see [7]).

Lemma 1 is proved using the probabilistic method. The next two lemmas about the distribution of splits for random domain elements may already be known; see e.g. [19, 7] for proofs of some closely related statements. We included proofs in appendices because we do not know a reference with proofs for exactly the statements needed here.

Lemma 2. *Assume $n \geq 1$. For $\mathbf{u}$ chosen uniformly at random from $\{0, ..., p-1\}^n$, for any $\mathbf{s} \in \{0, ..., p-1\}^n - \{\mathbf{0}\}$ for any $z \in \{0, ..., p-1\}$, we have*

$$\mathbf{Pr}(\mathbf{s} \cdot \mathbf{u} = z \mod p) = 1/p.$$

Proof: See Appendix A. □

Lemma 3. *Assume $n \geq 2$. For $\mathbf{u}$ chosen uniformly at random from $\{0, ..., p-1\}^n$, for any $\mathbf{s}, \mathbf{t} \in \{1, ..., p-1\}^n$ such that $\mathbf{s} \neq \mathbf{t}$, and for any $z \in \{0, ..., p-1\}$, we have*

$$\mathbf{Pr}(\mathbf{t} \cdot \mathbf{u} = z \mod p \mid \mathbf{s} \cdot \mathbf{u} = z \mod p) = 1/p.$$

Proof. See Appendix B. □

Armed with Lemmas 2 and 3, we are ready for the proof of Lemma 1.

Proof (of Lemma 1): Let S be an arbitrary subset of $\{1, ..., p-1\}^n$. Choose $\mathbf{u}$ uniformly at random from $\{0, ..., p-1\}^n$. For each $z \in \{0, ..., p-1\}$, let S_z be the (random) set of $\mathbf{s} \in S$ such that $\mathbf{s} \cdot \mathbf{u} = z \mod p$. Lemma 2 implies that, for all z,

$$\mathbf{E}(|S_z|) = |S|/p$$

and, since Lemmas 2 and 3 imply that the events that $\mathbf{s} \cdot \mathbf{u} = z$ are pairwise independent,

$$\mathbf{Var}(|S_z|) = \mathbf{Var}(1_{\mathbf{s} \cdot \mathbf{u} = z})|S| = (1/p)(1 - 1/p)|S| < |S|/p.$$

Using Chebyshev's inequality,

$$\mathbf{Pr}(|S_z| \geq |S|/p + 2\sqrt{|S|}) \leq \frac{1}{4p}.$$

Applying a union bound, with probability at least $3/4$,

$$\forall z, |S_z| \leq |S|/p + 2\sqrt{S},$$

completing the proof. □

Now we are ready for the learning lower bound.

Lemma 4.

$$\overline{\lim}_{n \to \infty} \overline{\lim}_{p \to \infty} \frac{\mathrm{opt}_{\mathrm{bandit}}(F_L(p, n))}{pn \ln p} \geq 1. \tag{5}$$

Proof: Choose $n \geq 3$ and $p \geq 5$. Consider an adversary that maintains a list F_t of members of

$$\{f_{\mathbf{a}} : \mathbf{a} \in \{1, ..., p - 1\}^n\} \subseteq F_L(p, n)$$

that are consistent with its previous answers, always answers "no", and picks $\mathbf{x}_t$ for round t that splits F_t as evenly as possible; that is, $\mathbf{x}_t$ minimizes the maximum, over potential values of $\hat{y}_t$, of $|F_t \cap \{f : f(\mathbf{x}_t) = \hat{y}_t\}|$. As long as $|F_t| \geq p^2 \ln p$, Lemma 1 implies that,

$$\begin{aligned} |F_{t+1}| &\geq |F_t| - \frac{|F_t|}{p} - 2\sqrt{|F_t|} \\ &\geq |F_t| - \frac{|F_t|}{p} - \frac{2|F_t|}{p\sqrt{\ln p}} \\ &= \left(1 - \left(\frac{1 + 2/\sqrt{\ln p}}{p}\right)\right)|F_t|. \end{aligned}$$

Thus, by induction, we have

$$|F_t| \geq \left(1 - \left(\frac{1 + 2/\sqrt{\ln p}}{p}\right)\right)^{t-1} (p - 1)^n.$$

The adversary can force m mistakes before $|F_t| < p^2 \ln p$ if

$$\left(1 - \frac{1 + 2/\sqrt{\ln p}}{p}\right)^{m-1} (p - 1)^n \geq p^2 \ln p$$

which is true for $m = (1 - o(1))np \ln p$, proving (5). □

4. Upper bound

The upper bound proof closely follows the arguments in [18, 4].

Lemma 5. *For any set F of functions from some set X to* $\{0, ..., k-1\}$*,*

$$\mathrm{opt}_{\mathrm{bandit}}(F) \leq (1 + o(1))(k \ln k)\mathrm{opt}_{\mathrm{std}}(F).$$

Proof: Consider an algorithm A_b for the bandit model, which uses an algorithm A_s for the standard model as a subroutine, defined as follows. Algorithm A_b maintains a list of copies of algorithm A_s that have been given different inputs. For $\alpha = \frac{1}{k \ln k}$, each copy of A_s is given a weight: if it has made m mistakes, its weight is α^m. In each round, A_b uses these weights to make its prediction by taking a weighted vote over the predictions made by the copies of A_s.

Algorithm A_b starts with a single copy. Whenever it makes a mistake, all copies of A_s that made a prediction that was not used by A_b "forget" the round – their state is rewound as if the round did not happen. Each copy of A_s that voted for the winner is cloned, including its state, to make $k-1$ copies, and each copy is given a different "guess" of $f(x_t)$.

Let W_t be the total weight of all of the copies of A_s before round t. Since one copy of A_s always gets correct information, for all t, we have

$$W_t \geq \alpha^{\mathrm{opt}_{\mathrm{std}}(F)}. \tag{6}$$

On the other hand, after each round t in which A_b makes a mistake, copies of A_s whose total weight is at least W_t/k are cloned to make $k-1$ copies, each with weight $\alpha < 1/(k-1)$ times its old weight. Thus

$$W_{t+1} \leq (1 - 1/k)W_t + (1/k)(\alpha(k-1)W_t) < (1 - 1/k)W_t + \alpha W_t$$

and, after A_b has made m mistakes,

$$W_t < (1 - 1/k + \alpha)^m < e^{-(1/k-\alpha)m}.$$

Combining with (6) yields

$$e^{-(1/k-\alpha)m} > \alpha^{\mathrm{opt}_{\mathrm{std}}(F)}$$

which implies $m \leq \frac{\ln(1/\alpha)\mathrm{opt}_{\mathrm{std}}(F)}{1/k-\alpha}$ and substituting the value of α completes the proof. □

5. Putting it together

Theorem 1 follows from (4), Lemma 4, and Lemma 5.

6. Two open problems

Can the analysis of $F_L(p, n)$ improve our understanding of the cost of bandit feedback in the agnostic case?

It is not hard to see that $\mathrm{opt}_{\mathrm{bs}}(k, 1) = k - 1 = \Theta(k)$, and the proofs of Lemmas 4 and 5 imply that $\mathrm{opt}_{\mathrm{bs}}(k, 3) = \Theta(k \log k)$. What about $\mathrm{opt}_{\mathrm{bs}}(k, 2)$?

Acknowledgements

We thank anonymous reviewers for helpful comments, and Yoram Singer for a valuable conversation.

Appendix A. Proof of Lemma 2

Pick i such that $s_i \neq 0$. We have

$$\begin{aligned}\mathbf{Pr}(\mathbf{u} \cdot \mathbf{s} = z \mod p) &= \mathbf{Pr}(u_i s_i = z - \sum_{j \neq i} u_j s_j \mod p)\\ &= \mathbf{Pr}(u_i = \left(z - \sum_{j \neq i} u_j s_j\right) s_i^{-1} \mod p)\\ &= 1/p,\end{aligned}$$

completing the proof.

Appendix B. Proof of Lemma 3

Let i be one component such that $s_i \neq t_i$. Let $\mathbf{s}'$, $\mathbf{t}'$ and $\mathbf{u}'$ be the projections of $\mathbf{s}$, $\mathbf{t}$ and $\mathbf{u}$ onto the indices other than i.

Lemma 2 implies that $\mathbf{s}' \cdot \mathbf{u}' \mod p$ is distributed uniformly on $\{0, ..., p-1\}$. Thus, after conditioning on the event that $\mathbf{s} \cdot \mathbf{u} = z \mod p$, u_i is uniform over $\{0, ..., p-1\}$, which implies

$$\begin{aligned}&\mathbf{Pr}(\mathbf{t} \cdot \mathbf{u} = z \mod p \mid \mathbf{s} \cdot \mathbf{u} = z \mod p)\\ &= \mathbf{Pr}(u_i(t_i - s_i) = (\mathbf{s}' - \mathbf{t}') \cdot \mathbf{u}' \mod p \mid \mathbf{s} \cdot \mathbf{u} = z \mod p)\\ &= 1/p,\end{aligned}$$

completing the proof.

References

[1] N. Abe, A. W. Biermann, and P. M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

[2] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[4] P. Auer and P. M. Long. Structural results about on-line learning models with and without queries. *Machine Learning*, 36(3):147–181, 1999.

[5] P. Auer, P. M. Long, W. Maass, and G. J. Woeginger. On the complexity of function learning. *Machine Learning*, 18(2-3):187–230, 1995.

[6] A. Blum. On-line algorithms in machine learning. In *Online algorithms*, pages 306–325. Springer, 1998.

[7] A. Blum. `https://www.cs.cmu.edu/~avrim/451f11/lectures/lect1004.pdf`, 2011. Accessed on Nov 20, 2016.

[8] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[9] J. L. Carter and M. N. Wegman. Universal classes of hash functions. In *Proceedings of the ninth annual ACM symposium on Theory of computing*, pages 106–112. ACM, 1977.

[10] W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, volume 15, pages 208–214, 2011.

[11] K. Crammer and C. Gentile. Multiclass classification with bandit feedback using adaptive regularization. *Machine learning*, 90(3):347–383, 2013.

[12] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

[13] A. Daniely and T. Halbertal. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, pages 93–104, 2013.

[14] E. Hazan and S. Kale. Newtron: an efficient bandit algorithm for online multiclass prediction. In *Advances in Neural Information Processing Systems*, pages 891–899, 2011.

[15] D. P. Helmbold, N. Littlestone, and P. M. Long. Apple tasting. *Information and Computation*, 161(2):85–139, 2000. Preliminary version in FOCS'92.

[16] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447. ACM, 2008.

[17] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[18] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. In *Foundations of Computer Science, 1989., 30th Annual Symposium on*, pages 256–261. IEEE, 1989.

[19] M. G. Luby and A. Wigderson. *Pairwise independence and derandomization*, volume 4. Now Publishers Inc, 2006.

[20] T. M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 1*, pages 305–310. Morgan Kaufmann Publishers Inc., 1977.

[21] C. R. Rao. Hypercubes of strength d leading to confounded designs in factorial experiments. *Bulletin of the Calcutta Mathematical Society*, 38:67–78, 1946.

[22] C. R. Rao. Factorial experiments derivable from combinatorial arrangements of arrays. *Supplement to the Journal of the Royal Statistical Society*, pages 128–139, 1947.

[23] H. Shvaytser. Linear manifolds are learnable from positive examples, 1988. Unpublished manuscript.