Bandit Multiclass Linear Classification: Efficient Algorithms for the Separable Case

Alina Beygelzimer[* 1] **Dávid Pál**[* 1] **Balázs Szörényi**[* 1] **Devanathan Thiruvenkatachari**[* 2] **Chen-Yu Wei**[* 3] **Chicheng Zhang**[* 4]

Abstract

We study the problem of efficient online multiclass linear classification with bandit feedback, where all examples belong to one of K classes and lie in the d-dimensional Euclidean space. Previous works have left open the challenge of designing efficient algorithms with finite mistake bounds when the data is linearly separable by a margin γ. In this work, we take a first step towards this problem. We consider two notions of linear separability, *strong* and *weak*.

1. Under the strong linear separability condition, we design an efficient algorithm that achieves a near-optimal mistake bound of $O\left(K/\gamma^2\right)$.
2. Under the more challenging weak linear separability condition, we design an efficient algorithm with a mistake bound of $\min(2^{\widetilde{O}(K \log^2(1/\gamma))}, 2^{\widetilde{O}(\sqrt{1/\gamma} \log K)})$.[1] Our algorithm is based on kernel Perceptron and is inspired by the work of Klivans & Servedio (2008) on improperly learning intersection of halfspaces.

1. Introduction

We study the problem of ONLINE MULTICLASS LINEAR CLASSIFICATION WITH BANDIT FEEDBACK (Kakade et al., 2008). The problem can be viewed as a repeated game between a learner and an adversary. At each time step t, the adversary chooses a labeled example (x_t, y_t) and reveals the feature vector x_t to the learner. Upon receiving x_t, the learner makes a prediction $\widehat{y}_t$ and receives feedback. In contrast with the standard full-information setting, where the feedback given is the correct label y_t, here the feedback is only a binary indicator of whether the prediction was correct or not. The protocol of the problem is formally stated below.

Protocol 1 ONLINE MULTICLASS LINEAR CLASSIFICATION WITH BANDIT FEEDBACK

Require: Number of classes K, number of rounds T.
Require: Inner product space $(V, \langle \cdot, \cdot \rangle)$.
for $t = 1, 2, \ldots, T$ **do**
  Adversary chooses example $(x_t, y_t) \in V \times \{1, 2, \ldots, K\}$ where x_t is revealed to the learner.
  Predict class label $\widehat{y}_t \in \{1, 2, \ldots, K\}$.
  Observe feedback $z_t = \mathbb{1}[\widehat{y}_t \neq y_t] \in \{0, 1\}$.

The performance of the learner is measured by its cumulative number of mistakes $\sum_{t=1}^T z_t = \sum_{t=1}^T \mathbb{1}[\widehat{y}_t \neq y_t]$, where $\mathbb{1}$ denotes the indicator function.

In this paper, we focus on the special case when the examples chosen by the adversary lie in $\mathbb{R}^d$ and are linearly separable with a margin. We introduce two notions of linear separability, *weak* and *strong*, formally stated in Definition 1. The standard notion of multiclass linear separability (Crammer & Singer, 2003) corresponds to the weak linear separability. For multiclass classification with K classes, weak linear separability requires that all examples from the same class lie in an intersection of $K-1$ halfspaces and all other examples lie in the complement of the intersection of the halfspaces. Strong linear separability means that examples from each class are separated from the remaining examples by a *single* hyperplane.

In the full-information feedback setting, it is well known (Crammer & Singer, 2003) that if all examples have norm at most R and are weakly linearly separable with a margin γ, then the MULTICLASS PERCEPTRON algorithm makes at most $\lfloor 2(R/\gamma)^2 \rfloor$ mistakes. It is also known that any (possibly randomized) algorithm must make $\frac{1}{2}\left\lfloor (R/\gamma)^2 \right\rfloor$ mistakes in the worst case. The MULTICLASS PERCEPTRON achieves an information-theoretically optimal mis-

[*]The authors are listed in alphabetical order. [1]Yahoo Research, New York, NY, USA [2]New York University, New York, NY, USA [3]University of Southern California, Los Angeles, CA, USA [4]Microsoft Research, New York, NY, USA. Correspondence to: Dávid Pál <davidko.pal@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1]We use the notation $\widetilde{O}(f(\cdot)) = O(f(\cdot) \operatorname{polylog}(f(\cdot)))$.

take bound, while being time and memory efficient.[2] [3]

The bandit feedback setting, however, is much more challenging. For the strongly linearly separable case, we are not aware of any prior efficient algorithm with a finite mistake bound. [4] We design a simple and efficient algorithm (Algorithm 1) that makes at most $O(K(R/\gamma)^2)$ mistakes in expectation. Its memory complexity and per-round time complexity are both $O(dK)$. The algorithm can be viewed as running K copies of the BINARY PERCEPTRON algorithm, one copy for each class. We prove that any (possibly randomized) algorithm must make $\Omega(K(R/\gamma)^2)$ mistakes in the worst case. The extra $O(K)$ multiplicative factor in the mistake bound, as compared to the full-information setting, is the price we pay for the bandit feedback, or more precisely, the lack of full-information feedback.

For the case when the examples are weakly linearly separable, it was open for a long time whether there exist *efficient* algorithms with finite mistake bound (Kakade et al., 2008; Beygelzimer et al., 2017). Furthermore, Kakade et al. (2008) ask the question: Is there *any* algorithm with a finite mistake bound that has no explicit dependence on the dimensionality of the feature vectors? We answer both questions affirmatively by providing an efficient algorithm with finite dimensionless mistake bound (Algorithm 2).[5]

The strategy used in Algorithm 2 is to construct a non-linear feature mapping ϕ and associated positive definite kernel $k(x, x')$ that makes the examples *strongly* linearly separable in a higher-dimensional space. We then use the kernelized version of Algorithm 1 for the strongly separable case. The kernel $k(x, x')$ corresponding to the feature mapping ϕ has a simple explicit formula and can be computed in $O(d)$ time, making Algorithm 2 computationally efficient. For details on kernel methods see e.g. (Schölkopf & Smola, 2002) or (Shawe-Taylor & Cristianini, 2004).

The number of mistakes of the kernelized algorithm depends on the margin in the corresponding feature space. We analyze how the mapping ϕ transforms the margin parameter of weak separability in the original space $\mathbb{R}^d$ into a margin parameter of strong separability in the new feature space. This problem is related to the problem of learning intersection of halfspaces and has been studied previously by Klivans & Servedio (2008). As a side result, we improve on the results of Klivans & Servedio (2008) by removing the dependency on the original dimension d.

The resulting kernelized algorithm runs in time polynomial in the original dimension of the feature vectors d, the number of classes K, and the number of rounds T. We prove that if the examples lie in the unit ball of $\mathbb{R}^d$ and are weakly linearly separable with margin γ, Algorithm 2 makes at most $\min(2^{\widetilde{O}(K \log^2(1/\gamma))}, 2^{\widetilde{O}(\sqrt{1/\gamma} \log K)})$ mistakes.

In Appendix G, we propose and analyze a very different algorithm for weakly linearly separable data. The algorithm is based on the obvious idea that two points that are close enough must have the same label.

Finally, we study two questions related to the computational and information-theoretic hardness of the problem. Any algorithm for the bandit setting collects information in the form of so called *strongly labeled* and *weakly labeled* examples. Strongly labeled examples are those for which we know the class label. Weakly labeled example is an example for which we know that class label can be anything except for one particular class. In Appendix H, we show that the offline problem of finding a multiclass linear classifier consistent with a set of strongly and weakly labeled examples is NP-hard. In Appendix I, we prove a lower bound on the number of mistakes of any algorithm that uses only strongly-labeled examples and ignores weakly labeled examples.

2. Related work

The problem of online bandit multiclass learning was initially formulated in the pioneering work of Auer & Long (1999) under the name of "weak reinforcement model". They showed that if all examples agree with some classifier from a prespecified hypothesis class $\mathcal{H}$, then the optimal mistake bound in the bandit setting can be upper bounded by the optimal mistake bound in the full information setting, times a factor of $(2.01 + o(1))K \ln K$. Long (2017) later improved the factor to $(1 + o(1))K \ln K$ and showed its near-optimality. Daniely & Helbertal (2013) extended the results to the setting where the performance of the algorithm is measured by its regret, i.e. the difference between the number of mistakes made by the algorithm and the number of mistakes made by the best classifier in $\mathcal{H}$ in hindsight. We remark that all algorithms developed in this context are computationally inefficient.

The linear classification version of this problem is initially studied by Kakade et al. (2008). They proposed two computationally inefficient algorithms that work in the weakly linearly separable setting, one with a mistake bound of $O(K^2 d \ln(d/\gamma))$, the other with a mistake bound of

[2] We call an algorithm computationally efficient, if its running time is polynomial in K, d, $1/\gamma$ and T.

[3] For completeness, we present these folklore results along with their proofs in Appendix A in the supplementary material.

[4] Although Chen et al. (2009) claimed that their Conservative OVA algorithm with PA-I update has a finite mistake bound under the strong linear separability condition, their Theorem 2 is incorrect: first, their Lemma 1 (with $C = +\infty$) along with their Theorem 1 implies a mistake upper bound of $(\frac{R}{\gamma})^2$, which contradicts the lower bound in our Theorem 3; second, their Lemma 1 cannot be directly applied to the bandit feedback setting.

[5] An inefficient algorithm was given by (Daniely & Helbertal, 2013).

$\widetilde{O}((K^2/\gamma^2)\ln T)$. The latter result was later improved by Daniely & Helbertal (2013), which gives a computationally inefficient algorithm with a mistake upper bound of $\widetilde{O}(K/\gamma^2)$. In addition, Kakade et al. (2008) propose the BANDITRON algorithm, a computationally efficient algorithm that has a $O(T^{2/3})$ regret against the multiclass hinge loss in the general setting, and has a $O(\sqrt{T})$ mistake bound in the γ-weakly linearly separable setting. In contrast to mild dependencies on the time horizon for mistake bounds of computationally inefficient algorithms, the polynomial dependence of BANDITRON's mistake bound on the time horizon is undesirable for problems with a long time horizon, in the weakly linearly separable setting. One key open question left by Kakade et al. (2008) is whether one can design computationally efficient algorithms that achieve mistake bounds that match or improve over those of inefficient algorithms. In this paper, we take a step towards answering this question, showing that efficient algorithms with mistake bounds quasipolynomial in $1/\gamma$ (for constant K) and quasipolynomial in K (for constant γ) can be obtained.

The general problem of linear bandit multiclass learning has received considerable attention (Abernethy & Rakhlin, 2009; Wang et al., 2010; Crammer & Gentile, 2013; Hazan & Kale, 2011; Beygelzimer et al., 2017; Foster et al., 2018). Chen et al. (2014); Zhang et al. (2018) study online bandit multiclass boosting under bandit feedback, where one can view boosting as linear classification by treating each base hypothesis as a separate feature. In the weakly linearly separable setting, however, these algorithms can only guarantee a mistake bound of $O(\sqrt{T})$ at best.

The problem considered here is a special case of the contextual bandit problem (Auer et al., 2003; Langford & Zhang, 2008). In this general problem, there is a hidden cost vector c_t associated with every prediction in round t. Upon receiving x_t and predicting $\widehat{y}_t \in \{1,\ldots,K\}$, the learner gets to observe the incurred cost $c_t(\widehat{y}_t)$. The goal of the learner is to minimize its regret with respect to the best predictor in some predefined policy class Π, given by $\sum_{t=1}^T c_t(\widehat{y}_t) - \min_{\pi\in\Pi}\sum_{t=1}^T c_t(\pi(x_t))$. Bandit multiclass learning is a special case where the cost $c_t(i)$ is the classification error $\mathbb{1}[i \neq y_t]$ and the policy class is the set of linear classifiers $\left\{x \mapsto \operatorname{argmax}_y (Wx)_y : W \in \mathbb{R}^{K\times d}\right\}$. There has been significant progress on the general contextual bandit problem assuming access to an optimization oracle that returns a policy in Π with the smallest total cost on any given set of cost-sensitive examples (Dudík et al., 2011; Agarwal et al., 2014; Rakhlin & Sridharan, 2016; Syrgkanis et al., 2016a;b). However, such an oracle abstracting efficient search through Π is generally not available in our setting due to computational hardness results (Arora et al., 1997).

Recently, Foster & Krishnamurthy (2018) developed a rich theory of contextual bandits with surrogate losses, focusing on regrets of the form $\sum_{t=1}^T c_t(\widehat{y}_t) - \min_{f\in\mathcal{F}}\sum_{t=1}^T \frac{1}{K}\sum_{i=1}^K c_t(i)\phi(f_i(x_t))$, where $\mathcal{F}$ contains score functions $f = (f_1,\ldots,f_K)$ such that $\sum_{i=1}^K f_i(\cdot) \equiv 0$, and $\phi(s) = \max(1-\frac{s}{\gamma},0)$ or $\min(1,\max(1-\frac{s}{\gamma},0))$. On one hand, it gives information-theoretic regret upper bounds for various settings of $\mathcal{F}$. On the other hand, it gives an efficient algorithm with an $O(\sqrt{T})$ regret against the benchmark of $\mathcal{F} = \left\{x \mapsto Wx : W \in \mathbb{R}^{K\times d}, \mathbb{1}^T W = 0\right\}$. A direct application of this result to ONLINE BANDIT MULTICLASS LINEAR CLASSIFICATION gives an algorithm with $O(\sqrt{T})$ mistake bound in the strongly linearly separable case.

3. Notions of linear separability

Let $[n] = \{1,2,\ldots,n\}$. We define two notions of linear separability for multiclass classification. The first notion is the standard notion of linear separability used in the proof of the mistake bound for the MULTICLASS PERCEPTRON algorithm (see e.g. Crammer & Singer, 2003). The second notion is stronger, i.e. more restrictive.

Definition 1 (Linear separability)**.** *Let $(V,\langle\cdot,\cdot\rangle)$ be an inner product space, K be a positive integer, and γ be a positive real number. We say that labeled examples $(x_1,y_1),(x_2,y_2),\ldots,(x_T,y_T) \in V\times[K]$ are*

weakly linearly separable with a margin γ *if there exist vectors $w_1,w_2,\ldots,w_K \in V$ such that*

$$\sum_{i=1}^K \|w_i\|^2 \le 1, \tag{1}$$

$$\langle x_t, w_{y_t}\rangle \ge \langle x_t, w_i\rangle + \gamma \quad \forall t\in[T]\ \forall i\in[K]\setminus\{y_t\}, \tag{2}$$

and strongly linearly separable with a margin γ *if there exist vectors $w_1,w_2,\ldots,w_K \in V$ such that*

$$\sum_{i=1}^K \|w_i\|^2 \le 1, \tag{3}$$

$$\langle x_t, w_{y_t}\rangle \ge \gamma/2 \quad \forall t\in[T], \tag{4}$$

$$\langle x_t, w_i\rangle \le -\gamma/2 \quad \forall t\in[T]\ \forall i\in[K]\setminus\{y_t\}. \tag{5}$$

The notion of strong linear separability has appeared in the literature; see e.g. (Chen et al., 2009). Intuitively, strong linear separability means that, for each class i, the set of examples belonging to class i and the set of examples belonging to the remaining $K-1$ classes are separated by a linear classifier w_i with margin $\frac{\gamma}{2}$.

It is easy to see that if a set of labeled examples is strongly linearly separable with margin γ, then it is also weakly linearly separable with the same margin (or larger). Indeed, if

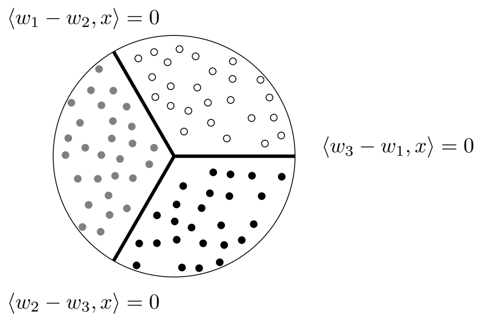


Figure 1. A set of labeled examples in $\mathbb{R}^2$. The examples belong to $K = 3$ classes colored white, gray and black respectively. Each class lies in a 120° wedge. In other words, each class lies in an intersection of two halfspaces. While the examples are weakly linearly separable with a positive margin γ, they are *not* strongly linearly separable with any positive margin γ. For instance, there does *not* exist a linear separator that separates the examples belonging to the gray class from the examples belonging to the remaining two classes.

$w_1, w_2, \dots, w_K \in V$ satisfy (3), (4), (5) then they satisfy (1) and (2).

In the special case of $K = 2$, if a set of labeled examples is weakly linearly separable with a margin γ, then it is also strongly linearly separable with the same margin. Indeed, if w_1, w_2 satisfy (1) and (2) then $w'_1 = \frac{w_1 - w_2}{2}$, $w'_2 = \frac{w_2 - w_1}{2}$ satisfy (3), (4), (5). Equation (3) follows from $\|w'_i\|^2 \le (\frac{1}{2}\|w_1\| + \frac{1}{2}\|w_2\|)^2 \le \frac{1}{2}\|w_1\|^2 + \frac{1}{2}\|w_2\|^2 \le \frac{1}{2}$ for $i = 1, 2$. Equations (4) and (5) follow from the fact that $w'_1 - w'_2 = w_1 - w_2$.

However, for any $K \ge 3$ and any inner product space of dimension at least 2, there exists a set of labeled examples that is weakly linearly separable with a positive margin γ but is not strongly linearly separable with any positive margin. Figure 1 shows one such set of labeled examples.

4. Algorithm for strongly linearly separable data

In this section, we consider the case when the examples are strongly linearly separable. We present an algorithm for this setting (Algorithm 1) and give an upper bound on its number of mistakes, stated as Theorem 2 below. The proof of the theorem can be found in Appendix B.

The idea behind Algorithm 1 is to use K copies of the BINARY PERCEPTRON algorithm, one copy per class; see e.g. (Shalev-Shwartz, 2012, Section 3.3.1). Upon seeing each example x_t, copy i predicts whether or not x_t belongs to class i. Multiclass predictions are done by evaluating all K binary predictors and outputting any class with a positive prediction. If all binary predictions are negative, the algorithm chooses a prediction uniformly at random from $\{1, 2, \dots, K\}$.

Algorithm 1 BANDIT ALGORITHM FOR STRONGLY LINEARLY SEPARABLE EXAMPLES

```
Require: Number of classes K, number of rounds T.
Require: Inner product space (V, ⟨·,·⟩).
Initialize w_1^(1) = w_2^(1) = ··· = w_K^(1) = 0
for t = 1, 2, ..., T do
    Observe feature vector x_t ∈ V
    Compute S_t = { i : 1 ≤ i ≤ K, ⟨w_i^(t), x_t⟩ ≥ 0 }
    if S_t = ∅ then
        Predict ŷ_t ~ Uniform({1, 2, ..., K})
        Observe feedback z_t = 1[ŷ_t ≠ y_t]
        if z_t = 1 then
            Set w_i^(t+1) = w_i^(t), ∀i ∈ {1, 2, ..., K}
        else
            Set w_i^(t+1) = w_i^(t), ∀i ∈ {1, 2, ..., K} \ {ŷ_t}
            Update w_{ŷ_t}^(t+1) = w_{ŷ_t}^(t) + x_t
    else
        Predict ŷ_t ∈ S_t chosen arbitrarily
        Observe feedback z_t = 1[ŷ_t ≠ y_t]
        if z_t = 1 then
            Set w_i^(t+1) = w_i^(t), ∀i ∈ {1, 2, ..., K} \ {ŷ_t}
            Update w_{ŷ_t}^(t+1) = w_{ŷ_t}^(t) − x_t
        else
            Set w_i^(t+1) = w_i^(t), ∀i ∈ {1, 2, ..., K}
```

Theorem 2 (Mistake upper bound). *Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space, K be a positive integer, γ be a positive real number, R be a non-negative real number. If the examples $(x_1, y_1), \dots, (x_T, y_T) \in V \times \{1, 2, \dots, K\}$ are strongly linearly separable with margin γ and $\|x_1\|, \|x_2\|, \dots, \|x_T\| \le R$ then the expected number of mistakes that Algorithm 1 makes is at most $(K-1)\lfloor 4(R/\gamma)^2 \rfloor$.*

The upper bound $(K-1)\lfloor 4(R/\gamma)^2 \rfloor$ on the expected number of mistakes of Algorithm 1 is optimal up to a constant factor, as long as the number of classes K is at most $O((R/\gamma)^2)$. This lower bound is stated as Theorem 3 below. The proof of the theorem can be found in Appendix B. Daniely & Helbertal (2013) provide a lower bound under the assumption of weak linear separability, which does not immediately imply a lower bound under the stronger notion.

Theorem 3 (Mistake lower bound). *Let γ be a positive real number, R be a non-negative real number and let $K \le (R/\gamma)^2$ be a positive integer. Any (possibly randomized) algorithm makes at least $((K-1)/2) \left\lfloor (R/\gamma)^2/4 \right\rfloor$ mistakes in expectation on some sequence of labeled examples $(x_1, y_1), (x_2, y_2), \dots, (x_T, y_T) \in V \times \{1, 2, \dots, K\}$*

for some inner product space $(V, \langle \cdot, \cdot \rangle)$ *such that the examples are strongly linearly separable with margin* γ *and satisfy* $\|x_1\|, \|x_2\|, \ldots, \|x_T\| \leq R$.

Remark. If $\gamma \leq R$ then, irrespective of any other conditions on K, R, and γ, a trivial lower bound on the expected number of mistakes of any randomized algorithm is $(K-1)/2$. To see this, note that the adversary can choose an example (Re_1, y), where e_1 is some arbitrary unit vector in V and y is a label chosen uniformly from $\{1, 2, \ldots, K\}$, and show this example K times. The sequence of examples trivially satisfies the strong linear separability condition, and the $(K-1)/2$ expected mistake lower bound follows from (Daniely & Helbertal, 2013, Claim 2).

Algorithm 1 can be extended to nonlinear classification using *positive definite kernels* (or *kernels*, for short), which are functions of the form $k : X \times X \to \mathbb{R}$ for some set X such that the matrix $\big(k(x_i, x_j)\big)_{i,j=1}^{m}$ is a symmetric positive semidefinite for any positive integer m and $x_1, x_2, \ldots, x_m \in X$ (Schölkopf & Smola, 2002, Definition 2.5).[6] As opposed to explicitly maintaining the weight vector for each class, the algorithm maintains the set of example-scalar pairs corresponding to the updates of the non-kernelized algorithm. As a direct consequence of Theorem 2 we get a mistake bound for the kernelized algorithm.

Theorem 4 (Mistake upper bound for kernelized algorithm)**.** *Let X be a non-empty set, let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. Let $\phi : X \to V$ be a feature map and let $k : X \times X \to \mathbb{R}$, $k(x, x') = \big\langle \phi(x), \phi(x') \big\rangle$ be the associated positive definite kernel. Let K be a positive integer, γ be a positive real number, R be a non-negative real number. If $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T) \in X \times \{1, 2, \ldots, K\}$ are labeled examples such that:*

1. *the mapped examples $(\phi(x_1), y_1), \ldots, (\phi(x_T), y_T)$ are strongly linearly separable with margin γ,*
2. $k(x_1, x_1), k(x_2, x_2), \ldots, k(x_T, x_T) \leq R^2$,

then the expected number of mistakes that Algorithm 2 makes is at most $(K-1)\lfloor 4(R/\gamma)^2 \rfloor$.

5. From weak separability to strong separability

In this section, we consider the case when the examples are weakly linearly separable. Throughout this section, we assume without loss of generality that all examples lie in the unit ball $\mathrm{B}(\mathbf{0}, 1) \subseteq \mathbb{R}^d$.[7] Note that Algorithm 1 alone does not guarantee a finite mistake bound in this setting, as weak linear separability does not imply strong linear separability.

Algorithm 2 KERNELIZED BANDIT ALGORITHM

Require: Number of classes K, number of rounds T.
Require: Kernel function $k(\cdot, \cdot)$.
Initialize $J_1^{(1)} = J_2^{(1)} = \cdots = J_K^{(1)} = \emptyset$
for $t = 1, 2, \ldots, T$ **do**
  Observe feature vector x_t.
  Compute
  $S_t = \Big\{ i \,:\, 1 \leq i \leq K,\ \sum_{(x,y) \in J_i^{(t)}} y k(x, x_t) \geq 0 \Big\}$
  if $S_t = \emptyset$ **then**
    Predict $\widehat{y}_t \sim \text{Uniform}(\{1, 2, \ldots, K\})$
    Observe feedback $z_t = \mathbb{1}\left[\widehat{y}_t \neq y_t\right]$
    if $z_t = 1$ **then**
      Set $J_i^{(t+1)} = J_i^{(t)}$ for all $i \in \{1, 2, \ldots, K\}$
    else
      Set $J_i^{(t+1)} = J_i^{(t)}, \forall i \in \{1, 2, \ldots, K\} \setminus \{\widehat{y}_t\}$
      Update $J_{\widehat{y}_t}^{(t+1)} = J_{\widehat{y}_t}^{(t)} \cup \big\{(x_t, +1)\big\}$
  else
    Predict $\widehat{y}_t \in S_t$ chosen arbitrarily
    Observe feedback $z_t = \mathbb{1}\left[\widehat{y}_t \neq y_t\right]$
    if $z_t = 1$ **then**
      Set $J_i^{(t+1)} = J_i^{(t)}, \forall i \in \{1, 2, \ldots, K\} \setminus \{\widehat{y}_t\}$
      Update $J_{\widehat{y}_t}^{(t+1)} = J_{\widehat{y}_t}^{(t)} \cup \big\{(x_t, -1)\big\}$
    else
      Set $J_i^{(t+1)} = J_i^{(t)}$ for all $i \in \{1, 2, \ldots, K\}$

We use a positive definite kernel function $k(\cdot, \cdot)$, namely a *rational kernel* (Shalev-Shwartz et al., 2011) whose corresponding feature map $\phi(\cdot)$ transforms any sequence of *weakly* linearly separable examples to a *strongly* linearly separable sequence of examples. Specifically, ϕ has the property that if a set of labeled examples in $\mathrm{B}(\mathbf{0}, 1)$ is weakly linearly separable with a margin γ, then after applying ϕ the examples become strongly linearly separable with a margin γ' and their squared norms are bounded by 2.[8] The parameter γ' is a function of the old margin γ and the number of classes K, and is specified in Theorem 5 below.

[6] For every kernel there exists an associated feature map $\phi : X \to V$ into some inner product space $(V, \langle \cdot, \cdot \rangle)$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$.

[7] Instead of working with feature vector x_t we can work with normalized feature vectors $\widehat{x}_t = \frac{x_t}{\|x_t\|}$. It can be easily checked that if $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ are weakly linearly separable with margin γ and $\|x_t\| \leq R$ for all t, then the normalized examples $(\widehat{x}_1, y_1), (\widehat{x}_2, y_2), \ldots, (\widehat{x}_T, y_T)$ are weakly linearly separable with margin γ/R.

[8] Other kernels, such as the polynomial kernel $k(x, x') = (1 + \langle x, x' \rangle)^d$, or the multinomial kernel (Goel & Klivans, 2017) $k(x, x') = \sum_{i=0}^{d} (\langle x, x' \rangle)^i$, will have similar properties for large enough d.

The rational kernel $k : \mathrm{B}(\mathbf{0},1) \times \mathrm{B}(\mathbf{0},1) \to \mathbb{R}$ is defined as

$$k(x,x') = \frac{1}{1 - \frac{1}{2}\left\langle x, x' \right\rangle_{\mathbb{R}^d}} . \tag{6}$$

Note that $k(x,x')$ can be evaluated in $O(d)$ time.

Consider the classical real separable Hilbert space $\ell_2 = \{x \in \mathbb{R}^\infty \ : \ \sum_{i=1}^\infty x_i^2 < +\infty\}$ equipped with the standard inner product $\left\langle x, x' \right\rangle_{\ell_2} = \sum_{i=1}^\infty x_i x_i'$. If we index the coordinates of ℓ_2 by d-tuples $(\alpha_1, \alpha_2, \ldots, \alpha_d)$ of non-negative integers, the feature map that corresponds to k is $\phi : \mathrm{B}(\mathbf{0},1) \to \ell_2$,

$$\big(\phi(x_1, x_2, \ldots, x_d)\big)_{(\alpha_1,\alpha_2,\ldots,\alpha_d)} = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d} \cdot \sqrt{2^{-(\alpha_1+\alpha_2+\cdots+\alpha_d)} \binom{\alpha_1+\alpha_2+\cdots+\alpha_d}{\alpha_1,\alpha_2,\ldots,\alpha_d}} \tag{7}$$

where $\binom{\alpha_1+\alpha_2+\cdots+\alpha_d}{\alpha_1,\alpha_2,\ldots,\alpha_d} = \frac{(\alpha_1+\alpha_2+\cdots+\alpha_d)!}{\alpha_1!\alpha_2!\ldots\alpha_d!}$ is the multinomial coefficient. It can be easily checked that

$$k(x,x') = \left\langle \phi(x), \phi(x') \right\rangle_{\ell_2} .$$

The last equality together with the formula for k implies that $k(x,x) < +\infty$ for any x in $\mathrm{B}(\mathbf{0},1)$ and thus in particular implies that $\phi(x)$ indeed lies in ℓ_2.

The following theorem is our main technical result in this section. We defer its proof to Section 5.1.

Theorem 5 (Margin transformation)**.** *Let* (x_1,y_1), (x_2,y_2), $\ldots$, $(x_T,y_T) \in \mathrm{B}(\mathbf{0},1) \times \{1,2,\ldots,K\}$ *be a sequence of labeled examples that is weakly linearly separable with margin* $\gamma > 0$*. Let* ϕ *be as defined in equation* (7) *and let*

$$\gamma_1 = \frac{\left[376\lceil\log_2(2K-2)\rceil \cdot \left\lceil\sqrt{\frac{2}{\gamma}}\right\rceil\right]^{\frac{-\lceil\log_2(2K-2)\rceil\cdot\lceil\sqrt{2/\gamma}\rceil}{2}}}{2\sqrt{K}},$$

$$\gamma_2 = \frac{\left(2^{s+1} r(K-1)(4s+2)\right)^{-(s+1/2)r(K-1)}}{4\sqrt{K}(4K-5)2^{K-1}} ,$$

where $r = 2\left\lceil\frac{1}{4}\log_2(4K-3)\right\rceil + 1$ *and* $s = \lceil\log_2(2/\gamma)\rceil$*. Then, the sequence of labeled examples transformed by* ϕ*, namely* $(\phi(x_1),y_1), (\phi(x_2),y_2), \ldots, (\phi(x_T),y_T)$*, is strongly linearly separable with margin* $\gamma' = \max\{\gamma_1,\gamma_2\}$*. In addition, for all* t *in* $\{1,\ldots,T\}$*,* $k(x_t,x_t) \le 2$*.*

Using this theorem we derive a mistake bound for Algorithm 2 with kernel (6) under the weak linear separability assumption.

Corollary 6 (Mistake upper bound)**.** *Let* K *be a positive integer and let* γ *be a positive real number. If* $(x_1,y_1),(x_2,y_2),\ldots,(x_T,y_T) \in \mathrm{B}(\mathbf{0},1) \times \{1,2,\ldots,K\}$ *is a sequence of weakly separable labeled examples with margin* $\gamma > 0$*, then the expected number of mistakes made by Algorithm 2 with kernel* $k(x,x')$ *defined by* (6) *is at most* $\min(2^{\widetilde{O}(K\log^2(1/\gamma))}, 2^{\widetilde{O}(\sqrt{1/\gamma}\log K)})$.

This corollary follows directly from Theorems 4 and 5. We remark that under the weakly linearly separable setting, (Daniely & Helbertal, 2013) gives a mistake lower bound of $\Omega(\frac{K}{\gamma^2})$ for *any algorithm* (see also Theorem 3). We leave the possibility of designing efficient algorithms that have mistakes bounds matching this lower bound as an important open question.

5.1. Proof of Theorem 5

Overview. The idea behind the construction and analysis of the mapping ϕ is polynomial approximation. Specifically, we construct K multivariate polynomials $p_1, p_2, \ldots, p_K$ such that

$$\forall t \in \{1,2,\ldots,T\}, \qquad p_{y_t}(x_t) \ge \frac{\gamma'}{2} , \tag{8}$$

$$\forall t \in \{1,2,\ldots,T\} \ \forall i \in \{1,2,\ldots,K\} \setminus \{y_t\}, \quad p_i(x_t) \le -\frac{\gamma'}{2} . \tag{9}$$

We then show (Lemma 9) that each polynomial p_i can be expressed as $\left\langle c_i, \phi(x) \right\rangle_{\ell_2}$ for some $c_i \in \ell_2$. This immediately implies that the examples $(\phi(x_1),y_1),\ldots,(\phi(x_T),y_T)$ are strongly linearly separable with a positive margin.

The conditions (8) and (9) are equivalent to that

$$\forall t \in \{1,2,\ldots,T\}, y_t = i \quad \Rightarrow \quad p_i(x_t) \ge \frac{\gamma'}{2} , \tag{10}$$

$$\forall t \in \{1,2,\ldots,T\}, y_t \ne i \quad \Rightarrow \quad p_i(x_t) \le -\frac{\gamma'}{2} . \tag{11}$$

hold for all $i \in \{1,2,\ldots,K\}$. We can thus fix i and focus on construction of one particular polynomial p_i.

Since examples $(x_1,y_1),(x_2,y_2),\ldots,(x_T,y_T)$ are weakly linearly separable, all examples from class i lie in

$$R_i^+ = \bigcap_{j \in \{1,2,\ldots,K\}\setminus\{i\}} \left\{ x \in \mathrm{B}(\mathbf{0},1) \ : \ \left\langle w_i^* - w_j^*, x \right\rangle \ge \gamma \right\},$$

and all examples from the remaining classes lie in

$$R_i^- = \bigcup_{j \in \{1,2,\ldots,K\}\setminus\{i\}} \left\{ x \in \mathrm{B}(\mathbf{0},1) \ : \ \left\langle w_i^* - w_j^*, x \right\rangle \le -\gamma \right\}.$$

Therefore, to satisfy conditions (10) and (11), it suffices to construct p_i such that

$$x \in R_i^+ \quad \Longrightarrow \quad p_i(x) \geq \frac{\gamma'}{2} \,, \tag{12}$$

$$x \in R_i^- \quad \Longrightarrow \quad p_i(x) \leq -\frac{\gamma'}{2} \,. \tag{13}$$

According to the well known Stone-Weierstrass theorem (see e.g. Davidson & Donsig, 2010, Section 10.10), on a compact set, multivariate polynomials uniformly approximate any continuous function. Roughly speaking, the conditions (12) and (13) mean that p_i approximates on $\mathrm{B}(\mathbf{0},1)$ a scalar multiple of the indicator function of the intersection of $K-1$ halfspaces $\bigcap_{j \in \{1,2,\dots,K\} \setminus \{i\}} \left\{ x : \left\langle w_i^* - w_j^*, x \right\rangle \geq 0 \right\}$ while within margin γ along the decision boundary, the polynomial is allowed to attain arbitrary values. It is thus clear such a polynomial exists.

We give two explicit constructions for such polynomial in Theorems 7 and 8. Our constructions are based on Klivans & Servedio (2008) which in turn uses the constructions from Beigel et al. (1995). More importantly, the theorems quantify certain parameters of the polynomial, which allows us to upper bound the transformed margin γ'.

Before we state the theorems, recall that a polynomial of d variables is a function $p : \mathbb{R}^d \to \mathbb{R}$ of the form

$$\begin{aligned} p(x) &= p(x_1, x_2, \dots, x_d) \\ &= \sum_{\alpha_1, \alpha_2, \dots, \alpha_d} c_{\alpha_1, \alpha_2, \dots, \alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} \end{aligned}$$

where the sum ranges over a finite set of d-tuples $(\alpha_1, \alpha_2, \dots, \alpha_d)$ of non-negative integers and $c_{\alpha_1, \alpha_2, \dots, \alpha_d}$'s are real coefficients. The *degree* of a polynomial p, denoted by $\deg(p)$, is the largest value of $\alpha_1 + \alpha_2 + \cdots + \alpha_d$ for which the coefficient $c_{\alpha_1, \alpha_2, \dots, \alpha_d}$ is non-zero. Following the terminology of Klivans & Servedio (2008), the *norm of a polynomial* p is defined as

$$\|p\| = \sqrt{\sum_{\alpha_1, \alpha_2, \dots, \alpha_d} \left(c_{\alpha_1, \alpha_2, \dots, \alpha_d} \right)^2} \,.$$

It is easy see that this is indeed a norm, since we can interpret it as the Euclidean norm of the vector of the coefficients of the polynomial.

Theorem 7 (Polynomial approximation of intersection of halfspaces I)**.** *Let $v_1, v_2, \dots, v_m \in \mathbb{R}^d$ be vectors such that $\|v_1\|, \|v_2\|, \dots, \|v_m\| \leq 1$. Let $\gamma \in (0,1)$. There exists a multivariate polynomial $p : \mathbb{R}^d \to \mathbb{R}$ such that*

1. *$p(x) \geq 1/2$ for all $x \in R^+ = \bigcap_{i=1}^m \{x \in \mathrm{B}(\mathbf{0},1) \,:\, \langle v_i, x \rangle \geq \gamma\}$,*
2. *$p(x) \leq -1/2$ for all $x \in R^- = \bigcup_{i=1}^m \{x \in \mathrm{B}(\mathbf{0},1) \,:\, \langle v_i, x \rangle \leq -\gamma\}$,*
3. *$\deg(p) = \lceil \log_2(2m) \rceil \cdot \left\lceil \sqrt{1/\gamma} \right\rceil$,*
4. *$\|p\| \leq \left[188 \lceil \log_2(2m) \rceil \cdot \left\lceil \sqrt{1/\gamma} \right\rceil \right]^{\frac{\lceil \log_2(2m) \rceil \cdot \lceil \sqrt{1/\gamma} \rceil}{2}}$.*

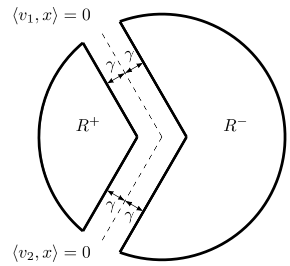


Figure 2. The figure shows the two regions R^+ and R^- used in parts 1 and 2 of Theorems 7 and 8 for the case $m = d = 2$ and a particular choice of vectors v_1, v_2 and margin parameter γ. The separating hyperplanes $\langle v_1, x \rangle = 0$ and $\langle v_2, x \rangle = 0$ are shown as dashed lines.

Theorem 8 (Polynomial approximation of intersection of halfspaces II)**.** *Let $v_1, v_2, \dots, v_m \in \mathbb{R}^d$ be vectors such that $\|v_1\|, \|v_2\|, \dots, \|v_m\| \leq 1$. Let $\gamma \in (0,1)$. Define*

$$r = 2 \left\lceil \frac{1}{4} \log_2(4m+1) \right\rceil + 1 \quad \text{and} \quad s = \lceil \log_2(1/\gamma) \rceil \,.$$

Then, there exists a multivariate polynomial $p : \mathbb{R}^d \to \mathbb{R}$ such that

1. *$p(x) \geq 1/2$ for all $x \in R^+ = \bigcap_{i=1}^m \{x \in \mathrm{B}(\mathbf{0},1) \,:\, \langle v_i, x \rangle \geq \gamma\}$,*
2. *$p(x) \leq -1/2$ for all $x \in R^- = \bigcup_{i=1}^m \{x \in \mathrm{B}(\mathbf{0},1) \,:\, \langle v_i, x \rangle \leq -\gamma\}$,*
3. *$\deg(p) \leq (2s+1)rm$,*
4. *$\|p\| \leq (4m-1)2^m \cdot \left(2^s rm(4s+2)\right)^{(s+1/2)rm}$.*

The proofs of the theorems are in Appendix D. The geometric interpretation of the two regions R^+ and R^- in the theorems is explained in Figure 2. Similar but weaker results were proved by Klivans & Servedio (2008). Specifically, our bounds in parts 1, 2, 3, 4 of Theorems 7 and 8 are independent of the dimension d.

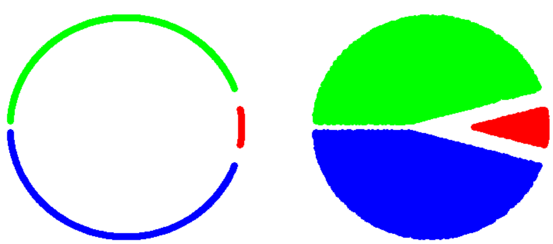
The following lemma establishes a correspondence between any multivariate polynomial in $\mathbb{R}^d$ and an element in ℓ_2, and gives an upper bound on its norm. Its proof follows from simple algebra, which we defer to Appendix C.

Lemma 9 (Norm bound)**.** *Let $p : \mathbb{R}^d \to \mathbb{R}$ be a multivariate polynomial. There exists $c \in \ell_2$ such that $p(x) = \langle c, \phi(x) \rangle_{\ell_2}$ and $\|c\|_{\ell_2} \leq 2^{\deg(p)/2}\|p\|$.*

Using the lemma and the polynomial approximation theorems, we can prove that the mapping ϕ maps any set of weakly linearly separable examples to a strongly linearly separable set of examples. Due to space constraints, we defer the full proof of Theorem 5 to Appendix E.

6. Experiments

In this section, we provide an empirical evaluation on our algorithms, verifying their effectiveness on linearly separable datasets. We generated strongly and weakly linearly separable datasets with $K = 3$ classes in $\mathbb{R}^3$ i.i.d. from two data distributions. Figures 3a and 3b show visualizations of the two datasets, along with detailed descriptions of the distributions.



(a) Strongly separable case (b) Weakly separable case

Figure 3. Strongly and weakly linearly separable datasets in $\mathbb{R}^3$ with $K = 3$ classes and $T = 5 \times 10^6$ examples. Here we show projections of the examples onto their first two coordinates, which lie in the ball of radius $1/\sqrt{2}$ centered at the origin. The third coordinate is $1/\sqrt{2}$ for all examples. Class 1 is depicted red. Classes 2 and 3 are depicted green and blue, respectively. 80% of the examples belong to class 1, 10% belong to class 2 and 10% belong to class 3. Class 1 lies in the angle interval $[-15°, 15°]$, while classes 2 and 3 lie in the angle intervals $[15°, 180°]$ and $[-180°, -15°]$ respectively. The examples are strongly and weakly linearly separable with a margin of $\gamma = 0.05$, respectively. (Examples lying within margin γ of the linear separators were rejected during sampling.)

We implemented Algorithm 1, Algorithm 2 with rational kernel (6) and used implementation of BANDITRON algorithm by Orabona (2009). We evaluated these algorithms on the two datasets. BANDITRON has an exploration rate parameter, for which we tried values $0.02, 0.01, 0.005, 0.002, 0.001, 0.0005$. Since all three algorithms are randomized, we run each algorithm 20 times. The average cumulative number of mistakes up to round t as a function of t are shown in Figures 4 and 5.

We can see that there is a tradeoff in the setting of the exploration rate for BANDITRON. With large exploration parameter, BANDITRON suffers from over-exploration, whereas with small exploration parameter, its model cannot be updated quickly enough. As expected, Algorithm 1 has a small number of mistakes in the strongly linearly separable setting, while having a large number of mistakes in the weakly linearly separable setting, due to the limited representation power of linear classifiers. In contrast, Algorithm 2 with rational kernel has a small number of mistakes in both settings, exhibiting strong adaptivity guarantees. Appendix F shows the decision boundaries that each of the algorithms learns by the end of the last round.

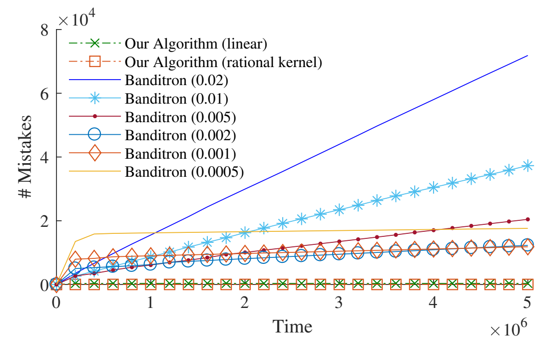


Figure 4. The average cumulative number of mistakes versus the number of rounds on the strongly linearly separable dataset in Figure 3a.

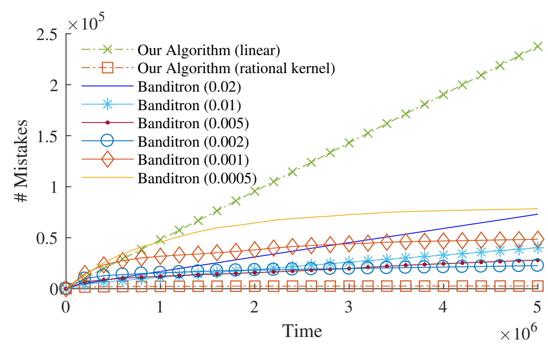


Figure 5. The average cumulative number of mistakes versus the number of rounds on the weakly linearly separable dataset in Figure 3b.

Acknowledgments

We thank Francesco Orabona and Wen Sun for helpful initial discussions, and thank Adam Klivans and Rocco Servedio for helpful discussions on (Klivans & Servedio, 2008) and pointing out the reference (Klivans & Servedio, 2004).

We also thank Dylan Foster, Akshay Krishnamurthy, and Haipeng Luo for providing a candidate solution to our problem.

References

Abernethy, J. and Rakhlin, A. An efficient bandit algorithm for $\sqrt{T}$-regret in online multiclass prediction? In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009)*, 2009.

Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. Taming the monster: a fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning (ICML 2014)*, 2014.

Arora, S., Babai, L., Stern, J., and Sweedyk, Z. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54(2):317–331, 1997.

Auer, P. and Long, P. M. Structural results about online learning models with and without queries. *Machine Learning*, 36(3):147–181, 1999.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003.

Beigel, R., Reingold, N., and Spielman, D. PP is closed under intersection. *Journal of Computer and System Sciences*, 50(2):191–202, 1995.

Beygelzimer, A., Orabona, F., and Zhang, C. Efficient online bandit multiclass learning with $\widetilde{O}(\sqrt{T})$ regret. In *International Conference on Machine Learning*, pp. 488–497, 2017.

Blum, A. L. and Rivest, R. L. Training a 3-node neural network is NP-complete. In *Machine learning: From theory to applications*, pp. 9–28. Springer, 1993.

Chen, G., Chen, G., Zhang, J., Chen, S., and Zhang, C. Beyond banditron: A conservative and efficient reduction for online multiclass prediction with bandit setting model. In *Ninth IEEE International Conference on Data Mining, 2009 (ICDM 2009)*, pp. 71–80. IEEE, 2009.

Chen, S.-T., Lin, H.-T., and Lu, C.-J. Boosting with online binary learners for the multiclass bandit problem. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 342–350, Bejing, China, 22–24 Jun 2014. PMLR.

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

Crammer, K. and Singer, Y. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3(Jan):951–991, 2003.

Daniely, A. and Helbertal, T. The price of bandit information in multiclass online classification. In *Conference on Learning Theory*, pp. 93–104, 2013.

Davidson, K. R. and Donsig, A. P. *Real analysis and Applications*. Springer, 2010.

Dudík, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., Reyzin, L., and Zhang, T. Efficient optimal learning for contextual bandits. In *UAI 2011*, pp. 169–178, 2011.

Foster, D. and Krishnamurthy, A. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. In *Advances in Neural Information Processing Systems*, 2018.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Logistic regression: The importance of being improper. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory (COLT 2018)*, volume 75 of *Proceedings of Machine Learning Research*, pp. 167–208. PMLR, 06–09 Jul 2018.

Garey, M. R. and Johnson, D. S. *Computers and intractability: A guide to the theory of NP-completeness*. Freeman, 1979.

Goel, S. and Klivans, A. Learning depth-three neural networks in polynomial time. *arXiv preprint arXiv:1709.06010*, 2017.

Hazan, E. and Kale, S. Newtron: An efficient bandit algorithm for online multiclass prediction. In *Advances in neural information processing systems*, pp. 891–899, 2011.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 440–447. ACM, 2008.

Klivans, A. R. and Servedio, R. A. Perceptron-like performance for learning intersections of halfspaces. *COLT*, Open problem, 2004.

Klivans, A. R. and Servedio, R. A. Learning intersections of halfspaces with a margin. *Journal of Computer and System Sciences*, 74(1):35–48, 2008.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS 20*, pp. 817–824, 2008.

Long, P. M. On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

Long, P. M. New bounds on the price of bandit feedback for mistake-bounded online multiclass learning. In *International Conference on Algorithmic Learning Theory*, pp. 3–10, 2017.

Mason, J. C. and Handscomb, D. C. *Chebyshev polynomials*. Chapman and Hall/CRC, 2002.

Orabona, F. *DOGMA: a MATLAB toolbox for Online Learning*, 2009. Software available at `http://dogma.sourceforge.net`.

Rakhlin, A. and Sridharan, K. BISTRO: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning (ICML 2016)*, pp. 1977–1985, 2016.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Shalev-Shwartz, S., Shamir, O., and Sridharan, K. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40(6):1623–1646, 2011.

Shawe-Taylor, J. and Cristianini, N. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Syrgkanis, V., Krishnamurthy, A., and Schapire, R. Efficient algorithms for adversarial contextual learning. In *ICML*, pp. 2159–2168, 2016a.

Syrgkanis, V., Luo, H., Krishnamurthy, A., and Schapire, R. E. Improved regret bounds for oracle-based adversarial contextual bandits. In *NIPS*, pp. 3135–3143, 2016b.

Wang, S., Jin, R., and Valizadegan, H. A potential-based framework for online multi-class learning with partial feedback. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 900–907, 2010.

Zhang, D., Jung, Y. H., and Tewari, A. Online multiclass boosting with bandit feedback. *arXiv preprint arXiv:1810.05290*, 2018.

A. Multiclass Perceptron

MULTICLASS PERCEPTRON is an algorithm for ONLINE MULTICLASS CLASSIFICATION. Both the protocol for the problem and the algorithm are stated below. The algorithm assumes that the feature vectors come from an inner product space $(V, \langle\cdot,\cdot\rangle)$.

Two results are folklore. The first result is Theorem 10 which states that if examples are linearly separable with margin γ and examples have norm at most R then the algorithm makes at most $\lfloor 2(R/\gamma)^2 \rfloor$ mistakes. The second result is Theorem 11 which states that under the same assumptions as in Theorem 11 *any* deterministic algorithm for ONLINE MULTICLASS CLASSIFICATION must make at least $\lfloor (R/\gamma)^2 \rfloor$ mistakes in the worst case.

Protocol 2 ONLINE MULTICLASS CLASSIFICATION

Require: Number of classes K, number of rounds T.
Require: Inner product space $(V, \langle\cdot,\cdot\rangle)$.
for $t = 1, 2, \ldots, T$ **do**
  Adversary chooses example $(x_t, y_t) \in V \times \{1, 2, \ldots, K\}$, where x_t is revealed to the learner.
  Predict class label $\widehat{y}_t \in \{1, 2, \ldots, K\}$.
  Observe feedback y_t.

Algorithm 3 MULTICLASS PERCEPTRON

Require: Number of classes K, number of rounds T.
Require: Inner product space $(V, \langle\cdot,\cdot\rangle)$.
Initialize $w_1^{(1)} = w_2^{(1)} = \cdots = w_K^{(1)} = 0$
for $t = 1, 2, \ldots, T$ **do**
  Observe feature vector $x_t \in V$
  Predict $\widehat{y}_t = \operatorname{argmax}_{i \in \{1,2,\ldots,K\}} \left\langle w_t^{(i)}, x_t \right\rangle$
  Observe $y_t \in \{1, 2, \ldots, K\}$
  if $\widehat{y}_t \neq y_t$ **then**
    Set $w_i^{(t+1)} = w_i^{(t)}$
    for all $i \in \{1, 2, \ldots, K\} \setminus \{y_t, \widehat{y}_t\}$
    Update $w_{y_t}^{(t+1)} = w_{y_t}^{(t)} + x_t$
    Update $w_{\widehat{y}_t}^{(t+1)} = w_{\widehat{y}_t}^{(t)} - x_t$
  else
    Set $w_i^{(t+1)} = w_i^{(t)}$ for all $i \in \{1, 2, \ldots, K\}$

Theorem 10 (Mistake upper bound (Crammer & Singer, 2003))**.** *Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space, let K be a positive integer, let γ be a positive real number and let R be a non-negative real number. If $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ is a sequence of labeled examples in $V \times \{1, 2, \ldots, K\}$ that are weakly linearly separable with margin γ and $\|x_1\|, \|x_2\|, \ldots, \|x_T\| \leq R$ then* MULTICLASS PERCEPTRON *algorithm makes at most $\lfloor 2(R/\gamma)^2 \rfloor$ mistakes.*

Proof. Let $M = \sum_{t=1}^{T} \mathbb{1}[\widehat{y}_t \neq y_t]$ be the number of mistakes the algorithm makes. Since the K-tuple $(w_1^{(t)}, w_2^{(t)}, \ldots, w_K^{(t)})$ changes only if a mistake is made, we can upper bound $\sum_{i=1}^{K} \left\| w_i^{(t)} \right\|^2$ in terms of number of mistakes. If a mistake happens in round t then

$$\begin{aligned}
\sum_{i=1}^{K} \left\| w_i^{(t+1)} \right\|^2 &= \left(\sum_{i \in \{1,2,\dots,K\} \setminus \{y_t, \widehat{y}_t\}} \left\| w_i^{(t)} \right\|^2 \right) + \left\| w_{y_t}^{(t)} + x_t \right\|^2 + \left\| w_{\widehat{y}_t}^{(t)} - x_t \right\|^2 \\
&= \left(\sum_{i \in \{1,2,\dots,K\} \setminus \{y_t, \widehat{y}_t\}} \left\| w_i^{(t)} \right\|^2 \right) + \left\| w_{y_t}^{(t)} \right\|^2 + \left\| w_{\widehat{y}_t}^{(t)} \right\|^2 + 2\|x_t\|^2 + 2 \left\langle w_{y_t}^{(t)} - w_{\widehat{y}_t}^{(t)}, x_t \right\rangle \\
&= \left(\sum_{i=1}^{K} \left\| w_i^{(t)} \right\|^2 \right) + 2\|x_t\|^2 + 2 \left\langle w_{y_t}^{(t)} - w_{\widehat{y}_t}^{(t)}, x_t \right\rangle \\
&\le \left(\sum_{i=1}^{K} \left\| w_i^{(t)} \right\|^2 \right) + 2\|x_t\|^2 \\
&\le \left(\sum_{i=1}^{K} \left\| w_i^{(t)} \right\|^2 \right) + 2R^2 \ .
\end{aligned}$$

So each time a mistake happens, $\sum_{i=1}^{K} \left\| w_i^{(t)} \right\|^2$ increases by at most $2R^2$. Thus,

$$\sum_{i=1}^{K} \left\| w_i^{(T+1)} \right\|^2 \le 2R^2 M \ . \tag{14}$$

Let $w_1^*, w_2^*, \dots, w_K^* \in V$ be vectors satisfying (1) and (2). We lower bound $\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(t)} \right\rangle$. This quantity changes only when a mistakes happens. If mistake happens in round t, we have

$$\begin{aligned}
\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(t+1)} \right\rangle &= \left(\sum_{i \in \{1,2,\dots,K\} \setminus \{y_t, \widehat{y}_t\}} \left\langle w_i^*, w_i^{(t)} \right\rangle \right) \\
&\quad + \left\langle w_{y_t}^*, w_{y_t}^{(t)} + x_t \right\rangle + \left\langle w_{\widehat{y}_t}^*, w_{\widehat{y}_t}^{(t)} - x_t \right\rangle \\
&= \left(\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(t)} \right\rangle \right) + \left\langle w_{y_t}^* - w_{\widehat{y}_t}^*, x_t \right\rangle \\
&\ge \left(\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(t)} \right\rangle \right) + \gamma \ .
\end{aligned}$$

Thus, after M mistakes,

$$\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(T+1)} \right\rangle \ge \gamma M \ .$$

We upper bound the left hand side by using Cauchy-Schwartz inequality twice and the condition (1) on $w_1^*, w_2^*, \dots, w_K^*$. We have

$$\begin{aligned}
\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(T+1)} \right\rangle &\le \sum_{i=1}^{K} \|w_i^*\| \cdot \left\| w_i^{(T+1)} \right\| \\
&\le \sqrt{\sum_{i=1}^{K} \|w_i^*\|^2} \sqrt{\sum_{i=1}^{K} \left\| w_i^{(T+1)} \right\|^2} \\
&\le \sqrt{\sum_{i=1}^{K} \left\| w_i^{(T+1)} \right\|^2} \ .
\end{aligned}$$

Combining the above inequality with Equations (14) and (A), we get

$$(\gamma M)^2 \le \sum_{i=1}^{K} \left\| w_i^{(T+1)} \right\|^2 \le 2R^2 M \ .$$

We conclude that $M \le 2(R/\gamma)^2$. Since M is an integer, $M \le \lfloor 2(R/\gamma)^2 \rfloor$. □

Theorem 11 (Mistake lower bound)**.** *Let K be a positive integer, let γ be a positive real number and let R be a non-negative real number. For any (possibly randomized) algorithm $\mathcal{A}$ for the* ONLINE MULTICLASS CLASSIFICATION *problem there exists an inner product space $(V, \langle \cdot, \cdot \rangle)$, a non-negative integer T and a sequence of labeled examples $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ examples in $V \times \{1, 2, \ldots, K\}$ that are weakly linearly separable with margin γ, the norms satisfy $\|x_1\|, \|x_2\|, \ldots, \|x_T\| \le R$ and the algorithm makes at least $\frac{1}{2}\lfloor (R/\gamma)^2 \rfloor$ mistakes.*

Proof. Let $T = \lfloor (R/\gamma)^2 \rfloor$, $V = \mathbb{R}^T$, and for all t in $\{1, \ldots, T\}$, define instance $x_t = Re_t$ where e_t is t-th element of the standard orthonormal basis of $\mathbb{R}^T$. Let labels $y_1, \ldots, y_T$ be chosen i.i.d uniformly at random from $\{1, 2, \ldots, K\}$ and independently of any randomness used by the algorithm $\mathcal{A}$.

We first show that the set of examples $(x_1, y_1), \ldots, (x_T, y_T)$ we have constructed is weakly linearly separable with margin γ. To prove that, we demonstrate vectors $w_1, w_2, \ldots, w_K$ satisfying conditions (1) and (2). We define

$$w_i = \frac{\gamma}{R} \sum_{\substack{t : 1 \le t \le T \\ y_t = i}} e_t \qquad \text{for } i = 1, 2, \ldots, K.$$

Let $a_i = |\{t \ : \ 1 \le t \le T, \ y_t = i\}|$ be the number of occurrences of label i. It is easy to see that

$$\|w_i\|^2 = \frac{\gamma^2}{R^2} \sum_{\substack{t : 1 \le t \le T \\ y_t = i}} \|e_t\|^2 = \frac{a_i \gamma^2}{R^2} \qquad \text{for } i = 1, 2, \ldots, K.$$

Since $\sum_{i=1}^K a_i = T$, $\sum_{i=1}^K \|w_i\|^2 = T \cdot \frac{\gamma^2}{R^2} \le 1$, i.e. the condition (1) holds. To verify condition (2) consider any labeled example (x_t, y_t). Then, for any i in $\{1, \ldots, K\}$, by the definition of w_i, we have

$$\begin{aligned} \langle w_i, x_t \rangle &= \frac{\gamma}{R} \sum_{\substack{s : 1 \le s \le T \\ y_s = i}} \langle e_s, Re_t \rangle \\ &= \gamma \cdot \sum_{\substack{s : 1 \le s \le T \\ y_s = i}} \mathbb{1}\left[s = t \right] \\ &= \gamma \cdot \mathbb{1}\left[y_t = i \right] \ . \end{aligned}$$

Therefore, if $i = y_t$, $\langle w_i, x_t \rangle = \gamma$; otherwise $i \neq y_t$, in which case $\langle w_i, x_t \rangle = 0$. Hence, condition (2) holds.

We now give a lower bound on the number of mistakes $\mathcal{A}$ makes. As y_t is chosen uniformly from $\{1, 2, \ldots, K\}$, independently from $\mathcal{A}$'s randomization and the first $t - 1$ examples,

$$\mathbf{E}[\mathbb{1}\left[\widehat{y}_t \neq y_t\right]] \ge 1 - \frac{1}{K} \ge \frac{1}{2} \ .$$

Summing over all t in $\{1, \ldots, T\}$, we conclude that

$$\mathbf{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\widehat{y}_t \neq y_t\right] \right] \ge \frac{T}{2} = \frac{1}{2} \lfloor (R/\gamma)^2 \rfloor,$$

which completes the proof. □

B. Proofs of Theorems 2 and 3

Proof of Theorem 2. Let $M = \sum_{t=1}^{T} z_t$ be the number of mistakes Algorithm 1 makes. Let $A = \sum_{t=1}^{T} \mathbb{1}\left[S_t \neq \emptyset\right] z_t$ be the number of mistakes in the rounds when $S_t \neq \emptyset$, i.e. the number of rounds line 18 is executed. In addition, let $B = \sum_{t=1}^{T} \mathbb{1}\left[S_t = \emptyset\right] z_t$ be the number of mistakes in the rounds when $S_t = \emptyset$. It can be easily seen that $M = A + B$.

Let $C = \sum_{t=1}^{T} \mathbb{1}\left[S_t = \emptyset\right](1 - z_t)$ be the number of rounds line 12 gets executed. Let $U = \sum_{t=1}^{T}(\mathbb{1}\left[S_t \neq \emptyset\right] z_t + \mathbb{1}\left[S_t = \emptyset\right](1 - z_t))$ be the number of rounds line 12 or 18 gets executed. In other words, U is the number of times the K-tuple of vectors $(w_1^{(t)}, w_2^{(t)}, \dots, w_K^{(t)})$ gets updated. It can be easily seen that $U = A + C$.

The key observation is that $\mathbf{E}[B] = (K-1)\,\mathbf{E}[C]$. To see this, note that if $S_t = \emptyset$, there is $1/K$ probability that the algorithm guesses the correct label ($z_t = 0$) and with probability $(K-1)/K$ algorithm's guess is incorrect ($z_t = 1$). Therefore,

$$\mathbf{E}[z_t | S_t = \emptyset] = \frac{K-1}{K},$$

$$\mathbf{E}[B] = \frac{K-1}{K}\,\mathbf{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[S_t = \emptyset\right]\right],$$

$$\mathbf{E}[C] = \frac{1}{K}\,\mathbf{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[S_t = \emptyset\right]\right].$$

Putting all the information together, we get that

$$\begin{aligned}\mathbf{E}[M] &= \mathbf{E}[A] + \mathbf{E}[B] \\ &= \mathbf{E}[A] + (K-1)\,\mathbf{E}[C] \\ &\le (K-1)\,\mathbf{E}[A + C] \\ &= (K-1)\,\mathbf{E}[U]\,. \end{aligned} \tag{15}$$

To finish the proof, we need to upper bound the number of updates U. We claim that $U \le \lfloor 4(R/\gamma)^2 \rfloor$ with probability 1. The proof of this upper bound is similar to the proof of the mistake bound for MULTICLASS PERCEPTRON algorithm. Let $w_1^*, w_2^*, \dots, w_K^* \in V$ be vectors that satisfy (3), (4) and (5). The K-tuple $(w_1^{(t)}, w_2^{(t)}, \dots, w_K^{(t)})$ changes only if there is an update in round t. We investigate how $\sum_{i=1}^{K} \left\|w_i^{(t)}\right\|^2$ and $\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(t)} \right\rangle$ change. If there is an update in round t, by lines 12 and 18, we always have $w_{\widehat{y}_t}^{(t+1)} = w_{\widehat{y}_t}^{(t)} + (-1)^{z_t} x_t$, and for all $i \neq \widehat{y}_t$, $w_i^{(t+1)} = w_i^{(t)}$. Therefore,

$$\begin{aligned}\sum_{i=1}^{K} \left\|w_i^{(t+1)}\right\|^2 &= \left(\sum_{i \in \{1,2,\dots,K\} \setminus \{\widehat{y}_t\}} \left\|w_i^{(t)}\right\|^2\right) + \left\|w_{\widehat{y}_t}^{(t+1)}\right\|^2 \\ &= \left(\sum_{i \in \{1,2,\dots,K\} \setminus \{\widehat{y}_t\}} \left\|w_i^{(t)}\right\|^2\right) + \left\|w_{\widehat{y}_t}^{(t)} + (-1)^{z_t} x_t\right\|^2 \\ &= \left(\sum_{i=1}^{K} \left\|w_i^{(t)}\right\|^2\right) + \|x_t\|^2 + \underbrace{(-1)^{z_t} 2 \left\langle w_{\widehat{y}_t}^{(t)}, x_t \right\rangle}_{\le 0} \\ &\le \left(\sum_{i=1}^{K} \left\|w_i^{(t)}\right\|^2\right) + \|x_t\|^2 \\ &\le \left(\sum_{i=1}^{K} \left\|w_i^{(t)}\right\|^2\right) + R^2\,. \end{aligned}$$

The inequality that $(-1)^{z_t} 2 \left\langle w_{\widehat{y}_t}^{(t)}, x_t \right\rangle \le 0$ is from a case analysis: if line 12 is executed, then $z_t = 0$ and $\left\langle w_{\widehat{y}_t}^{(t)}, x_t \right\rangle < 0$; otherwise line 18 is executed, in which case $z_t = 1$ and $\left\langle w_{\widehat{y}_t}^{(t)}, x_t \right\rangle \ge 0$.

Hence, after U updates,

$$\sum_{i=1}^{K} \left\| w_i^{(T+1)} \right\|^2 \le R^2 U \; . \tag{16}$$

Similarly, if there is an update in round t, we have

$$\begin{aligned}
\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(t)} \right\rangle &= \left(\sum_{i \in \{1,2,\ldots,K\} \setminus \{\widehat{y}_t\}} \left\langle w_i^*, w_i^{(t)} \right\rangle \right) + \left\langle w_{\widehat{y}_t}^*, w_{\widehat{y}_t}^{(t+1)} \right\rangle \\
&= \left(\sum_{i \in \{1,2,\ldots,K\} \setminus \{\widehat{y}_t\}} \left\langle w_i^*, w_i^{(t)} \right\rangle \right) + \left\langle w_{\widehat{y}_t}^*, w_{\widehat{y}_t}^{(t)} + (-1)^{z_t} x_t \right\rangle \\
&= \left(\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(t)} \right\rangle \right) + (-1)^{z_t} \left\langle w_{\widehat{y}_t}^*, x_t \right\rangle \\
&\ge \left(\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(t)} \right\rangle \right) + \frac{\gamma}{2},
\end{aligned}$$

where the last inequality follows from a case analysis on z_t and Definition 1: if $z_t = 0$, then $\widehat{y}_t = y_t$, by Equation (4), we have that $\left\langle w_{\widehat{y}_t}^*, x_t \right\rangle \ge \frac{\gamma}{2}$; if $z_t = 1$, then $\widehat{y}_t \ne y_t$, by Equation (5), we have that $\left\langle w_{\widehat{y}_t}^*, x_t \right\rangle \le -\frac{\gamma}{2}$.

Thus, after U updates,

$$\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(T+1)} \right\rangle \ge \frac{\gamma U}{2} \; . \tag{17}$$

Applying Cauchy-Schwartz's inequality twice, and using assumption (3), we get that

$$\begin{aligned}
\sum_{i=1}^{K} \left\langle w_i^*, w_i^{(T+1)} \right\rangle &\le \sum_{i=1}^{K} \|w_i^*\| \cdot \left\| w_i^{(T+1)} \right\| \\
&\le \sqrt{\sum_{i=1}^{K} \|w_i^*\|^2} \sqrt{\sum_{i=1}^{K} \left\| w_i^{(T+1)} \right\|^2} \\
&\le \sqrt{\sum_{i=1}^{K} \left\| w_i^{(T+1)} \right\|^2} \; .
\end{aligned}$$

Combining the above inequality with Equations (16) and (17), we get

$$\left(\frac{\gamma U}{2} \right)^2 \le \sum_{i=1}^{K} \left\| w_i^{(T+1)} \right\|^2 \le R^2 U \; .$$

We conclude that $U \le 4(R/\gamma)^2$. Since U is an integer, $U \le \lfloor 4(R/\gamma)^2 \rfloor$.

Applying Equation (15), we get

$$\mathbf{E}[M] \le (K-1)\,\mathbf{E}[U] \le (K-1) \lfloor 4(R/\gamma)^2 \rfloor \; . \quad \square$$

Proof of Theorem 3. Let $M = \left\lfloor \frac{1}{4}(R/\gamma)^2 \right\rfloor$. Let $V = \mathbb{R}^{M+1}$ equipped with the standard inner product. Let $e_1, e_2, \ldots, e_{M+1}$ be the standard orthonormal basis of V. We define vectors $v_1, v_2, \ldots, v_M \in V$ where $v_j = \frac{R}{\sqrt{2}}(e_j +$

$e_{M+1})$ for $j = 1, 2, \ldots, M$. Let $\ell_1, \ell_2, \ldots, \ell_M$ be chosen i.i.d. uniformly at random from $\{1, 2, \ldots, K\}$ and independently of any randomness used the by algorithm $\mathcal{A}$. Let $T = M(K-1)$. We define examples $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ as follows. For any $j = 1, 2, \ldots, M$ and any $h = 1, 2, \ldots, K-1$,

$$(x_{(j-1)(K-1)+h}, y_{(j-1)(K-1)+h}) = (v_j, \ell_j)$$

The norm of each example is exactly R. The examples are strongly linearly separable with margin γ. To see that, consider $w_1^*, w_2^*, \ldots, w_K^* \in V$ defined by

$$w_i^* = \sqrt{2}\frac{\gamma}{R}\left(\sum_{j\,:\,\ell_j=i} e_j\right) - \frac{\sqrt{2}}{2}\frac{\gamma}{R}e_{M+1}$$

for $i = 1, 2, \ldots, K$.

For $i \in \{1, 2, \ldots, K\}$ and $j \in \{1, 2, \ldots, M\}$, consider the inner product of w_i^* and v_j. If $i = \ell_j$, $\langle w_i^*, v_j \rangle = \gamma - \frac{\gamma}{2} = \frac{\gamma}{2}$; otherwise $i \neq \ell_j$, in which case $\langle w_i^*, v_j \rangle = 0 - \frac{\gamma}{2} = -\frac{\gamma}{2}$. This means that $w_1^*, w_2^*, \ldots, w_K^*$ satisfy conditions (4) and (5). Condition (3) is satisfied since

$$\begin{aligned}
\sum_{i=1}^{K} \|w_i^*\|^2 &= 2\frac{\gamma^2}{R^2}\sum_{j=1}^{M}\|e_j\|^2 + \frac{\gamma^2}{2R^2}K\|e_{M+1}\|^2 \\
&= 2\frac{\gamma^2}{R^2}M + \frac{\gamma^2}{2R^2}K \le \frac{1}{2} + \frac{1}{2} = 1 \,.
\end{aligned}$$

It remains to lower bound the expected number of mistakes of $\mathcal{A}$. For any $j \in \{1, 2, \ldots, M\}$, consider the expected number of mistakes the algorithm makes in rounds $(K-1)(j-1)+1, (K-1)(j-1)+2, \ldots, (K-1)j$.

Define a filtration of σ-algebras $\{\mathcal{B}_j\}_{j=0}^{M}$, where $\mathcal{B}_j = \sigma((x_1, y_1, \hat{y}_1), \ldots, (x_{(K-1)j}, y_{(K-1)j}, \hat{y}_{(K-1)j}))$ for every j in $\{1, 2, \ldots, M\}$. By Claim 2 of Daniely & Helbertal (2013), as ℓ_j is chosen uniformly from $\{1, \ldots, K\}$ and independent of $\mathcal{B}_{j-1}$ and $\mathcal{A}$'s randomness,

$$\mathbf{E}\left[\sum_{t=(K-1)(j-1)+1}^{(K-1)j} z_t \,\middle|\, \mathcal{B}_{j-1}\right] \ge \frac{K-1}{2} \,.$$

This implies that

$$\mathbf{E}\left[\sum_{t=(K-1)(j-1)+1}^{(K-1)j} z_t\right] \ge \frac{K-1}{2} \,.$$

Summing over all j in $\{1, 2, \ldots, M\}$,

$$\mathbf{E}\left[\sum_{t=1}^{(K-1)M} z_t\right] \ge \frac{K-1}{2} \cdot M = \frac{K-1}{2}\left\lfloor \frac{1}{4}(R/\gamma)^2 \right\rfloor \,.$$

Thus there exists a particular sequence of examples for which the algorithm makes at least $\frac{K-1}{2}\left\lfloor \frac{1}{4}(R/\gamma)^2 \right\rfloor$ mistakes in expectation over its internal randomization. $\square$

C. Proof of Lemma 9

Proof. Note that the polynomial p can be written as $p(x) = \sum_{\alpha_1, \alpha_2, \ldots, \alpha_d} c'_{\alpha_1, \alpha_2, \ldots, \alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}$. We define $c \in \ell_2$ using the multi-index notation as

$$c_{\alpha_1, \alpha_2, \ldots, \alpha_d} = \frac{c'_{\alpha_1, \alpha_2, \ldots, \alpha_d} 2^{(\alpha_1+\alpha_2+\cdots+\alpha_d)/2}}{\sqrt{\binom{\alpha_1+\alpha_2+\cdots+\alpha_d}{\alpha_1, \alpha_2, \ldots, \alpha_d}}}$$

for all tuples $(\alpha_1, \alpha_2, \ldots, \alpha_d)$ such that $\alpha_1 + \alpha_2 + \cdots + \alpha_d \le \deg(p)$. Otherwise, we define $c_{\alpha_1, \alpha_2, \ldots, \alpha_d} = 0$. By the definition of ϕ, $\langle c, \phi(x) \rangle_{\ell_2} = p(x)$.

Whether $\alpha_1 + \ldots + \alpha_d \le \deg(p)$, we always have:

$$|c_{\alpha_1,\alpha_2,\ldots,\alpha_d}| \le 2^{(\alpha_1+\alpha_2+\cdots+\alpha_d)/2}|c'_{\alpha_1,\alpha_2,\ldots,\alpha_d}| \le 2^{\deg(p)/2}|c'_{\alpha_1,\alpha_2,\ldots,\alpha_d}| \, .$$

Therefore,

$$\|c\|_{\ell_2} \le 2^{\deg(p)/2}\sqrt{\sum_{\alpha_1,\alpha_2,\ldots,\alpha_d}(c'_{\alpha_1,\alpha_2,\ldots,\alpha_d})^2} = 2^{\deg(p)/2}\|p\| \, . \qquad \square$$

D. Proofs of Theorems 7 and 8

In this section, we follow the construction of Klivans & Servedio (2008) (which in turn uses the constructions of Beigel et al. (1995)) to establish two polynomials of low norm, such that it takes large positive values in

$$\bigcap_{i=1}^{m}\left\{x \in \mathbb{R}^d \, : \, \|x\| \le 1, \, \langle v_i, x\rangle \ge \gamma\right\}$$

and takes large negative values in

$$\bigcup_{i=1}^{m}\left\{x \in \mathbb{R}^d \, : \, \|x\| \le 1, \, \langle v_i, x\rangle \le -\gamma\right\} .$$

We improve the norm bound analysis of Klivans & Servedio (2008) in two aspects:

1. Our upper bounds on the norm of the polynomials do not have any dependency on the dimensionality d.
2. We remove the requirement that the fractional part of input x must be above some threshold in Theorem 8.

A lot of the proof details are similar to those of Klivans & Servedio (2008); nevertheless, we provide a self-contained full proof here.

For the proofs of the theorems we need several auxiliary results.

Lemma 12 (Simple inequality)**.** *For any real numbers $b_1, b_2, \ldots, b_n$,*

$$\left(\sum_{i=1}^{n} b_i\right)^2 \le n\sum_{i=1}^{n} b_i^2 \, .$$

Proof. The lemma follows from Cauchy-Schwartz inequality applied to vectors $(b_1, b_2, \ldots, b_n)$ and $(1, 1, \ldots, 1)$. $\square$

Lemma 13 (Bound on binomial coefficients)**.** *For any integers n, k such that $n \ge k \ge 0$,*

$$\binom{n}{k} \le (n-k+1)^k \, .$$

Proof. If $k = 0$, the inequality trivially holds. For the rest of the proof we can assume $k \ge 1$. We write the binomial coefficient as

$$\begin{aligned}\binom{n}{k} &= \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} \\ &= \frac{n}{k}\cdot\frac{n-1}{k-1}\cdots\frac{n-k+1}{1} \, .\end{aligned}$$

We claim that

$$\frac{n}{k} \le \frac{n-1}{k-1} \le \cdots \le \frac{n-k+1}{1}$$

from which the lemma follows by upper bounding all the fractions by $n-k+1$. It remains to prove that for any $j = 0, 1, \ldots, k-1$,

$$\frac{n-j+1}{k-j+1} \le \frac{n-j}{k-j} \ .$$

Multiplying by the (positive) denominators, we get an equivalent inequality

$$(n-j+1)(k-j) \le (n-j)(k-j+1) \ .$$

We multiply out the terms and get

$$nk - kj + k - nj + j^2 - j \le nk - nj + n - kj + j^2 - j \ .$$

We cancel common terms and get an equivalent inequality $k \ge n$, which holds by the assumption. $\square$

Lemma 14 (Properties of the norm of polynomials)**.**

1. *Let $p_1, p_2, \ldots, p_n$ be multivariate polynomials and let $p(x) = \prod_{j=1}^n p_j(x)$ be their product. Then, $\|p\|^2 \le n^{\sum_{j=1}^n \deg(p_j)} \prod_{j=1}^n \|p_j\|^2$.*

2. *Let q be a multivariate polynomial of degree at most s and let $p(x) = (q(x))^n$. Then, $\|p\|^2 \le n^{ns} \|q\|^{2n}$.*

3. *Let be $p_1, p_2, \ldots, p_n$ be multivariate polynomials. Then, $\left\|\sum_{j=1}^n p_j\right\| \le \sum_{j=1}^n \|p_j\|$. Consequently, $\left\|\sum_{j=1}^n p_j\right\|^2 \le n \sum_{j=1}^n \|p_j\|^2$.*

Proof. Using multi-index notation we can write any multivariate polynomial p as

$$p(x) = \sum_A c_A x^A$$

where $A = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a multi-index (i.e. a d-tuple of non-negative integers), $x^A = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d}$ is a monomial and $c_A = c_{\alpha_1, \alpha_2, \ldots, \alpha_d}$ is the corresponding real coefficient. The sum is over a finite subset of d-tuples of non-negative integers. Using this notation, the norm of a polynomial p can be written as

$$\|p\| = \sqrt{\sum_A (c_A)^2} \ .$$

For a multi-index $A = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ we define its 1-norm as $\|A\|_1 = \alpha_1 + \alpha_2 + \cdots + \alpha_d$.

To prove the part 1, we express p_j as

$$p_j(x) = \sum_{A_j} c_{A_j}^{(j)} x^{A_j} \ .$$

Since $p(x) = \prod_{i=1}^n p_j(x)$, the coefficients of its expansion $p(x) = \sum_A c_A x^A$ are

$$c_A = \sum_{\substack{(A_1, A_2, \ldots, A_n) \\ A_1 + A_2 + \cdots + A_n = A}} c_{A_1}^{(1)} c_{A_2}^{(2)} \cdots c_{A_n}^{(n)} \ .$$

Therefore,

$$
\begin{aligned}
\|p\|^2 &= \sum_A (c_A)^2 \\
&= \sum_A \left(\sum_{\substack{(A_1, A_2, \dots, A_n) \\ A_1 + A_2 + \dots + A_n = A}} c^{(1)}_{A_1} c^{(2)}_{A_2} \cdots c^{(n)}_{A_n} \right)^2 \\
&= \sum_A \left(\sum_{\substack{(A_1, A_2, \dots, A_n) \\ A_1 + A_2 + \dots + A_n = A}} \prod_{j=1}^n c^{(j)}_{A_j} \right)^2
\end{aligned}
$$

and

$$
\begin{aligned}
\prod_{i=1}^n \|p_i\|^2 &= \prod_{i=1}^n \left(\sum_{A_i} (c^{(i)}_{A_i})^2 \right) \\
&= \sum_{(A_1, A_2, \dots, A_n)} \prod_{j=1}^n (c^{(j)}_{A_j})^2 \\
&= \sum_{(A_1, A_2, \dots, A_n)} \left(\prod_{j=1}^n c^{(j)}_{A_j} \right)^2 \\
&= \sum_A \sum_{\substack{(A_1, A_2, \dots, A_n) \\ A_1 + A_2 + \dots + A_n = A}} \left(\prod_{j=1}^n c^{(j)}_{A_j} \right)^2
\end{aligned}
$$

where in both cases the outer sum is over multi-indices A such that $\|A\|_1 \le \deg(p)$. Lemma 12 implies that for any multi-index A,

$$
\left(\sum_{\substack{(A_1, A_2, \dots, A_n) \\ A_1 + A_2 + \dots + A_n = A}} \prod_{j=1}^n c^{(j)}_{A_j} \right)^2 \le M_A \sum_{\substack{(A_1, A_2, \dots, A_n) \\ A_1 + A_2 + \dots + A_n = A}} \left(\prod_{j=1}^n c^{(j)}_{A_j} \right)^2 .
$$

where M_A is the number of n-tuples $(A_1, A_2, \dots, A_n)$ such that $A_1 + A_2 + \dots + A_n = A$.

To finish the proof, it is sufficient to prove that $M_A \le n^{\deg(p)}$ for any A such that $\|A\|_1 \le \deg(p)$. To prove this inequality, consider a multi-index $A = (\alpha_1, \alpha_2, \dots, \alpha_d)$ and consider its i-th coordinate α_i. In order for $A_1 + A_2 + \dots + A_n = A$ to hold, the i-th coordinates of $A_1, A_2, \dots, A_n$ need to sum to α_i. There are exactly $\binom{\alpha_i + n - 1}{\alpha_i}$ possibilities for the choice of i-th coordinates of $A_1, A_2, \dots, A_n$. The total number of choices is thus

$$
M_A = \prod_{i=1}^d \binom{\alpha_i + n - 1}{\alpha_i} .
$$

Using Lemma 13, we upper bound it as

$$
M_A \le \prod_{i=1}^d n^{\alpha_i} = n^{\|A\|_1} \le n^{\deg(p)} .
$$

Part 2 follows from the part 1 by setting $p_1 = p_2 = \dots p_n = q$.

The first inequality of part 3 follows from triangle inequality in Euclidean spaces, by viewing the polynomials $p = \sum_A c_A x^A$ as multidimensional vectors (c_A), and $\|p\| = \|(c_A)\|$.

For the second inequality, by Lemma 12, we have

$$\left\|\sum_{j=1}^{n} p_j\right\|^2 = \left(\left\|\sum_{j=1}^{n} p_j\right\|\right)^2 \le \left(\sum_{j=1}^{n}\|p_j\|\right)^2 \le n\sum_{j=1}^{n}\|p_j\|^2 \ .$$

□

D.1. Proof of Theorem 7

To construct the polynomial p we use Chebyshev polynomials of the first kind. Chebyshev polynomials of the fist kind form an infinite sequence of polynomials $T_0(z), T_1(z), T_2(z), \ldots$ of single real variable z. They are defined by the recurrence

$$\begin{aligned} T_0(z) &= 1 \ , \\ T_1(z) &= z \ , \\ T_{n+1}(z) &= 2zT_n(z) - T_{n-1}(z), \quad \text{for } n \ge 1. \end{aligned}$$

Chebyshev polynomials have a lot of interesting properties. We will need properties listed in Proposition 15 below. Interested reader can learn more about Chebyshev polynomials from the book by Mason & Handscomb (2002).

Proposition 15 (Properties of Chebyshev polynomials)**.** *Chebyshev polynomials satisfy*

1. $\deg(T_n) = n$ *for all* $n \ge 0$.
2. *If* $n \ge 1$*, the leading coefficient of* $T_n(z)$ *is* 2^{n-1}.
3. $T_n(\cos(\theta)) = \cos(n\theta)$ *for all* $\theta \in \mathbb{R}$ *and all* $n \ge 0$.
4. $T_n(\cosh(\theta)) = \cosh(n\theta)$ *for all* $\theta \in \mathbb{R}$ *and all* $n \ge 0$.
5. $|T_n(z)| \le 1$ *for all* $z \in [-1, 1]$ *and all* $n \ge 0$.
6. $T_n(z) \ge 1 + n^2(z-1)$ *for all* $z \ge 1$ *and all* $n \ge 0$.
7. $\|T_n\| \le (1+\sqrt{2})^n$ *for all* $n \ge 0$

Proof of Proposition 15. The first two properties can be easily proven by induction on n using the recurrence.

We prove the third property by induction on n. Indeed, by definition

$$T_0(\cos(\theta)) = 1 = \cos(0\theta) \quad \text{and} \quad T_1(\cos(\theta)) = \cos(\theta) \ .$$

For $n \ge 1$, we have

$$\begin{aligned} T_{n+1}(\cos(\theta)) &= 2\cos(\theta)T_n(\cos(\theta)) - T_{n-1}(\cos(\theta)) \\ &= 2\cos(\theta)\cos(n\theta) - \cos((n-1)\theta)) \ , \end{aligned}$$

where the last step follow by induction hypothesis. It remains to show that the last expression equals $\cos((n+1)\theta)$. This can be derived from the trigonometric formula

$$\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta) \ .$$

By substituting $\alpha = n\theta$ and $\beta = \theta$, we get two equations

$$\begin{aligned} \cos((n+1)\theta) &= \cos(n\theta)\cos(\theta) - \sin(n\theta)\sin(\theta) \ , \\ \cos((n-1)\theta) &= \cos(n\theta)\cos(\theta) + \sin(n\theta)\sin(\theta) \ . \end{aligned}$$

Summing them yields

$$\cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos(n\theta)\cos(\theta)$$

which finishes the proof.

The fourth property has the similar proof as the third property. It suffices to replace $\cos$ and $\sin$ with $\cosh$ and $\sinh$ respectively.

The fifth property follows from the third property. Indeed, for any $z \in [-1,1]$ there exists $\theta \in \mathbb{R}$ such that $\cos\theta = z$. Thus, $|T_n(z)| = |T_n(\cos(\theta))| = |\cos(n\theta)| \leq 1$.

The sixth property is equivalent to

$$T_n(\cosh(\theta)) \geq 1 + n^2(\cosh(\theta) - 1) \qquad \text{for all } \theta \geq 0,$$

since $\cosh(\theta) = \frac{e^{\theta}+e^{-\theta}}{2}$ is an even continuous function that maps $\mathbb{R}$ onto $[1,+\infty)$, is strictly decreasing on $(-\infty,0]$, and is strictly increasing on $[0,\infty)$. Using the fourth property the last inequality is equivalent to

$$\cosh(n\theta) \geq 1 + n^2(\cosh(\theta) - 1) \qquad \text{for all } \theta \geq 0.$$

For $\theta = 0$, both sides are equal to 1. Thus, it is sufficient to prove that the derivative of the left hand side is greater or equal to the derivative of the right hand side. Recalling that $[\cosh(\theta)]' = \sinh(\theta)$, this means that we need to show that

$$\sinh(n\theta) \geq n\sinh(\theta) \qquad \text{for all } \theta \geq 0.$$

To prove this inequality we use the summation formula

$$\sinh(\alpha+\beta) = \sinh(\alpha)\cosh(\beta) + \sinh(\beta)\cosh(\beta) \ .$$

If α, β are non-negative then $\sinh(\alpha), \sinh(\beta)$ are non-negative and $\cosh(\alpha), \cosh(\beta) \geq 1$. Hence,

$$\sinh(\alpha+\beta) \geq \sinh(\alpha) + \sinh(\beta) \qquad \text{for any } \alpha, \beta \geq 0.$$

This implies that (using induction on n) that $\sinh(n\theta) \geq n\sinh(\theta)$ for all $\theta \geq 0$.

We verify the seventh property by induction on n. For $n = 0$ and $n = 1$ the inequality trivially holds, since $\|T_0\| = \|T_1\| = 1$. For $n \geq 1$, since $T_{n+1}(z) = 2zT_n(z) - T_{n-1}(z)$,

$$\begin{aligned}
\|T_{n+1}\| &\leq 2\|T_n\| + \|T_{n-1}\| \\
&\leq 2(1+\sqrt{2})^n + (1+\sqrt{2})^{n-1} \\
&= (1+\sqrt{2})^{n-1}(2(1+\sqrt{2})+1) \\
&= (1+\sqrt{2})^{n-1}(3+2\sqrt{2}) \\
&= (1+\sqrt{2})^{n-1}(1+\sqrt{2})^2 \\
&= (1+\sqrt{2})^{n+1} \ .
\end{aligned}$$

□

We are now ready to prove Theorem 7. Let $r = \lceil \log_2(2m) \rceil$ and $s = \left\lceil \sqrt{\frac{1}{\gamma}} \right\rceil$. We define the polynomial $p : \mathbb{R}^d \to \mathbb{R}$ as

$$p(x) = m + \frac{1}{2} - \sum_{i=1}^{m} \left(T_s(1 - \langle v_i, x \rangle) \right)^r \ .$$

It remains to show that p has properties 1–5.

To verify the first property notice that if $x \in \mathbb{R}^d$ satisfies $\|x\| \leq 1$ and $\langle v_i, x \rangle \geq \gamma$ then since $\|v_i\| \leq 1$ we have $\langle v_i, x \rangle \in [0,1]$. Thus, $T_s(1 - \langle v_i, x \rangle)$ and $\left(T_s(1 - \langle v_i, x \rangle)\right)^r$ lie in the interval $[-1,1]$. Therefore,

$$p(x) \geq m + \frac{1}{2} - m \geq \frac{1}{2} \ .$$

To verify the second property consider any $x \in \bigcup_{i=1}^m \left\{x \in \mathbb{R}^d \ : \|x\| \le 1, \ \langle v_i, x \rangle \le -\gamma \right\}$. Clearly, $\|x\| \le 1$ and there exists at least one $i \in \{1, 2, \dots, m\}$ such that $\langle v_i, x \rangle \le -\gamma$. Therefore, $1 - \langle v_i, x \rangle \ge 1 + \gamma$ and Proposition 15 (part 6) imply that

$$T_s(1 - \langle v_i, x \rangle) \ge 1 + s^2 \gamma \ge 2$$

and thus

$$\left(T_s(1 - \langle v_i, x \rangle)\right)^r \ge 2^r \ge 2m \ .$$

On the other hand for any $j \in \{1, 2, \dots, m\}$, we have $\left\langle v_j, x \right\rangle \in [-1, 1]$ and thus $1 - \left\langle v_j, x \right\rangle$ lies in the interval $[0, 2]$. According to Proposition 15 (parts 5 and 6), $T_s(1 - \left\langle v_j, x \right\rangle) \ge -1$. Therefore,

$$\begin{aligned} p(x) &= m + \frac{1}{2} - \left(T_s(1 - \langle v_i, x \rangle)\right)^r - \sum_{\substack{j \, : \, 1 \le j \le m \\ j \neq i}} \Big(T_s(1 - \left\langle v_j, x \right\rangle)\Big)^r \\ &\le m + \frac{1}{2} - 2m + (m-1) \le -\frac{1}{2} \ . \end{aligned}$$

The third property follows from the observation that the degree of p is the same as the degree of any one of the terms $\left(T_s(1 - \langle v_i, x \rangle)\right)^r$ which is $r \cdot s$.

To prove the fourth property, we need to upper bound the norm of p. Let $f_i(x) = 1 - \langle v_i, x \rangle$, let $g_i(x) = T_s(1 - \langle v_i, x \rangle)$ and let $h_i(x) = (T_s(1 - \langle v_i, x \rangle))^r$. We have

$$\|f_i\|^2 = 1 + \|v_i\|^2 \le 1 + 1 = 2 \ .$$

Let $T_s(z) = \sum_{j=0}^s c_j z^j$ be the expansion of s-th Chebyshev polynomial. Then,

$$\begin{aligned} \|g_i\|^2 &= \left\| \sum_{j=0}^s c_j (f_i)^j \right\|^2 \\ &\le (s+1) \sum_{j=0}^s \left\| c_j (f_i)^j \right\|^2 \qquad \text{(by part 3 of Lemma 14)} \\ &= (s+1) \sum_{j=0}^s (c_j)^2 \left\| (f_i)^j \right\|^2 \\ &\le (s+1) \sum_{j=0}^s (c_j)^2 j^j \|f_i\|^{2j} \qquad \text{(by part 2 of Lemma 14)} \\ &\le (s+1) \sum_{j=0}^s (c_j)^2 j^j 2^{2j} \\ &\le (s+1) s^s 2^{2s} \sum_{j=0}^s (c_j)^2 \\ &= (s+1) s^s 2^{2s} \|T_s\|^2 \\ &= (s+1) s^s 2^{2s} (1 + \sqrt{2})^{2s} \qquad \text{(by part 7 of Proposition 15)} \\ &= (s+1) \left(4(1+\sqrt{2})^2 s\right)^s \\ &\le \left(8(1+\sqrt{2})^2 s\right)^s \\ &\le (47s)^s \ . \end{aligned}$$

where we used that $s+1 \le 2^s$ for any non-negative integer s. Finally,

$$\begin{aligned}
\|p\| &\le m + \frac{1}{2} + \sum_{i=1}^m \left\|(g_i)^r\right\| \\
&= m + \frac{1}{2} + \sum_{i=1}^m \sqrt{\left\|(g_i)^r\right\|^2} \\
&\le m + \frac{1}{2} + \sum_{i=1}^m \sqrt{r^{rs}\|g_i\|^{2r}} \\
&\le m + \frac{1}{2} + m r^{rs/2} \left(47s\right)^{rs/2} \\
&= m + \frac{1}{2} + m \left(47rs\right)^{rs/2} \ .
\end{aligned}$$

We can further upper bound the last expression by using that $m \le \frac{1}{2}2^r$. Since $r, s \ge 1$,

$$\begin{aligned}
\|p\| &\le m + \frac{1}{2} + m \left(47rs\right)^{rs/2} \\
&\le \frac{1}{2}2^r + \frac{1}{2} + \frac{1}{2}2^r \left(47rs\right)^{rs/2} \\
&\le 2^r + \frac{1}{2}2^r \left(47rs\right)^{rs/2} \\
&= 2^r \left(1 + \frac{1}{2}\left(47rs\right)^{rs/2}\right) \\
&= 2^r \left(47rs\right)^{rs/2} \\
&\le 4^{rs/2} \left(47rs\right)^{rs/2} \\
&\le \left(188rs\right)^{rs/2} \ .
\end{aligned}$$

Substituting for r and s finishes the proof.

D.2. Proof of Theorem 8

We define several univariate polynomials

$$\begin{aligned}
P_n(z) &= (z-1)\prod_{i=1}^n (z-2^i)^2, \quad \text{for } n \ge 0, \\
A_{n,k}(z) &= (P_n(z))^k - (P_n(-z))^k, \quad \text{for } n, k \ge 0, \\
B_{n,k}(z) &= -(P_n(z))^k - (P_n(-z))^k, \quad \text{for } n, k \ge 0.
\end{aligned}$$

We define the polynomial $q : \mathbb{R}^d \to \mathbb{R}$ as

$$q(x) = \left[\sum_{i=1}^m A_{s,r}\left(\frac{\langle v_i, x\rangle}{\gamma}\right) \prod_{\substack{j \,:\, 1 \le j \le m \\ j \ne i}} B_{s,r}\left(\frac{\langle v_j, x\rangle}{\gamma}\right)\right] - \left(m - \frac{1}{2}\right)\prod_{j=1}^m B_{s,r}\left(\frac{\langle v_j, x\rangle}{\gamma}\right) \ .$$

Finally, we define $p(x) = 2^{-s(s+1)rm+1}q(x)$. We are going to show that this polynomial p satisfies the required properties.

For convenience we define univariate rational function

$$S_{n,k}(z) = \frac{A_{n,k}(z)}{B_{n,k}(z)}, \quad \text{for } n, k \ge 0,$$

and a multivariate rational function

$$Q(x) = \left(\sum_{i=1}^{m} S_{s,r} \left(\frac{\langle v_i, x \rangle}{\gamma} \right) \right) - \left(m - \frac{1}{2} \right) .$$

It is easy to verify that

$$q(x) = Q(x) \prod_{j=1}^{m} B_{s,r} \left(\frac{\langle v_j, x \rangle}{\gamma} \right) .$$

Lemma 16 (Properties of P_n)**.**

1. *If* $z \in [0, 1]$ *then* $P_n(-z) \leq P_n(z) \leq 0$.
2. *If* $z \in [1, 2^n]$ *then* $0 \leq 4P_n(z) \leq -P_n(-z)$.
3. *If* $z \geq 0$ *then* $-P_n(-z) \geq 2^{n(n+1)}$.

Proof. To prove the first part, note that $P_n(z)$ and $P_n(-z)$ are non-positive for $z \in [0, 1]$. We can write $\frac{P_n(z)}{P_n(-z)}$ as a product of $n + 1$ non-negative fractions

$$\frac{P_n(z)}{P_n(-z)} = \frac{1 - z}{1 + z} \prod_{i=1}^{n} \frac{(z + 2^i)^2}{(z - 2^i)^2} .$$

The first part follows from the observation that each fraction is upper bounded by 1.

To prove the second part, notice that $P_n(z)$ is non-negative and $P_n(-z)$ is non-positive for any $z \in [1, 2^n]$. Now, fix $z \in [1, 2^n]$ and let $j \in \{1, 2, \ldots, n\}$ be such that $2^{j-1} \leq z \leq 2^j$. This implies that $(z + 2^j)^2 \geq (2^j)^2 \geq 4(z - 2^j)^2$. We can write $\frac{P_n(z)}{-P_n(-z)}$ as a product of $n + 1$ non-negative fractions

$$\frac{P_n(z)}{-P_n(-z)} = \frac{z - 1}{z + 1} \cdot \frac{(z - 2^j)^2}{(z + 2^j)^2} \prod_{\substack{i \,:\, 1 \leq i \leq n \\ i \neq j}} \frac{(z - 2^i)^2}{(z + 2^i)^2} .$$

The second part follows from the observation that the second fraction is upper bounded by $1/4$ and all other fractions are upper bounded by 1.

The third part follows from

$$-P_n(-z) = (1 + z) \prod_{i=1}^{n} (z + 2^i)^2 \geq \prod_{i=1}^{n} 2^{2i} = 2^{n(n+1)} .$$

$\square$

Lemma 17 (Properties of $S_{n,r}$ and $B_{n,r}$)**.** *Let* n, m *be non-negative integers. Let* $r = 2 \left\lceil \frac{1}{4} \log_2(4m + 1) \right\rceil + 1$. *Then,*

1. *If* $z \in [1, 2^n]$ *then* $S_{n,r}(z) \in [1, 1 + \frac{1}{2m}]$.
2. *If* $z \in [-2^n, -1]$ *then* $S_{n,r}(z) \in [-1 - \frac{1}{2m}, -1]$.
3. *If* $z \in [-1, 1]$ *then* $|S_{n,r}(z)| \leq 1$.
4. *If* $z \in [-2^n, 2^n]$ *then* $B_{n,r}(z) \geq \left(1 - \frac{1}{4m+1} \right) 2^{n(n+1)r}$.

Proof. Note that $B_{n,r}(z)$ is an even function and $A_{n,r}(z)$ is an odd function. Therefore, $S_{n,r}(z)$ is odd. Also notice that r is an odd integer.

1. Observe that $S_{n,r}(z)$ can be written as

$$S_{n,r}(z) = \frac{1 + \left(-\frac{P_n(z)}{P_n(-z)}\right)^r}{1 - \left(-\frac{P_n(z)}{P_n(-z)}\right)^r} = \frac{1+c}{1-c}$$

where $c = \left(-\frac{P_n(z)}{P_n(-z)}\right)^r$. Since $z \in [1, 2^n]$, by part 2 of Lemma 16, $c \in [0, \frac{1}{4^r}]$. Since $r \geq \frac{1}{2}\log_2(4m+1)$, this means that $c \in [0, \frac{1}{4m+1}]$. Thus, $S_{n,r}(z) = \frac{1+c}{1-c} \in [1, 1 + \frac{1}{2m}]$.

2. Since $S_{n,r}(z)$ is odd, the statement follows from part 1.

3. Recall that $S_{n,r}(z)$ can be written as

$$S_{n,r}(z) = \frac{1+c}{1-c}$$

where $c = \left(-\frac{P_n(z)}{P_n(-z)}\right)^r$. If $z \in [0, 1]$, by part 1 of Lemma 16 and the fact that r is odd, $c \in [-1, 0]$, and thus, $S_{n,r}(z) = \frac{1+c}{1-c} \in [0, 1]$. Since $S_{n,r}(z)$ is odd, for $z \in [-1, 0]$, $S_{n,r}(z) \in [-1, 0]$.

4. Since $B_{n,r}(z)$ is even, we can without loss generality assume that $z \geq 0$. We consider two cases.
Case $z \in [0, 1]$. Since r is odd and $P_n(z)$ is non-positive,

$$\begin{aligned} B_{n,r}(z) &= -(P_n(z))^r + \left(-P_n(-z)\right)^r \\ &\geq \left(-P_n(-z)\right)^r \geq 2^{n(n+1)r} \\ &\geq 2^{n(n+1)r}\left(1 - \frac{1}{4m+1}\right) . \end{aligned}$$

where the second last inequality follows from part 3 of Lemma 16.
Case $z \in [1, 2^n]$. Since r is odd,

$$\begin{aligned} B_{n,r}(z) &= \left(-P_n(-z)\right)^r \left(1 - \left(-\frac{P_n(z)}{P_n(-z)}\right)^r\right) \\ &= \left(-P_n(-z)\right)^r (1-c) \end{aligned}$$

where $c = \left(-\frac{P_n(z)}{P_n(-z)}\right)^r$. Since $z \in [1, 2^n]$, by part 2 of Lemma 16, $c \in [0, \frac{1}{4^r}]$. By the definition of r that means that $c \in [0, \frac{1}{4m+1}]$. Thus,

$$\begin{aligned} B_{n,r}(z) &\geq \left(-P_n(-z)\right)^r \left(1 - \frac{1}{4m+1}\right) \\ &\geq 2^{n(n+1)r}\left(1 - \frac{1}{4m+1}\right) . \end{aligned}$$

where the last inequality follows from part 3 of Lemma 16.

$\square$

Lemma 18 (Properties of $Q(x)$)**.** *The rational function $Q(x)$ satisfies*

1. $Q(x) \geq \frac{1}{2}$ *for all* $x \in \bigcap_{i=1}^{m} \left\{x \in \mathbb{R}^d \,:\, \|x\| \leq 1,\ \langle v_i, x\rangle \geq \gamma\right\}$,

2. $Q(x) \leq -\frac{1}{2}$ *for all* $x \in \bigcup_{i=1}^{m} \left\{x \in \mathbb{R}^d \,:\, \|x\| \leq 1,\ \langle v_i, x\rangle \leq -\gamma\right\}$.

Proof. To prove part 1, consider any $x \in \bigcap_{i=1}^m \left\{x \in \mathbb{R}^d \, : \, \|x\| \le 1, \, \langle v_i, x \rangle \ge \gamma \right\}$. Then, $\frac{\langle v_i, x \rangle}{\gamma} \in [1, \frac{1}{\gamma}]$. By part 1 of Lemma 17, $S_{s,r}\left(\frac{\langle v_i,x\rangle}{\gamma}\right) \in [1, 1 + \frac{1}{2m}]$ and in particular $S_{s,r}\left(\frac{\langle v_i,x\rangle}{\gamma}\right) \ge 1$. Thus,

$$
\begin{aligned}
Q(x) &= \left(\sum_{i=1}^m S_{s,r}\left(\frac{\langle v_i, x\rangle}{\gamma}\right) \right) - (m - 1/2) \\
&\ge m - (m - 1/2) \\
&= 1/2 \, .
\end{aligned}
$$

To prove part 2, consider any $x \in \bigcup_{i=1}^m \left\{x \in \mathbb{R}^d \, : \, \|x\| \le 1, \, \langle v_i, x \rangle \le -\gamma \right\}$. Observe that $\frac{\langle v_i, x \rangle}{\gamma} \in [-\frac{1}{\gamma}, \frac{1}{\gamma}]$. Consider $S_{s,r}\left(\frac{\langle v_i,x\rangle}{\gamma}\right)$ for any $i \in \{1, 2, \dots, m\}$. Parts 1,2, and 3 of Lemma 17 and the fact $1/\gamma \le 2^s$ imply that $S_{s,r}\left(\frac{\langle v_i,x\rangle}{\gamma}\right) \le 1 + \frac{1}{2m}$ for all $i \in \{1, 2, \dots, m\}$. By the choice of x, there exists $j \in \{1, 2, \dots, m\}$ such that $\langle v_j, x \rangle \le -\gamma$. Part 2 of Lemma 17 implies that $S_{s,r}\left(\frac{\langle v_j,x\rangle}{\gamma}\right) \in [-1 - \frac{1}{2m}, -1]$. Thus,

$$
\begin{aligned}
Q(x) &= \left(\sum_{i=1}^m S_{s,r}\left(\frac{\langle v_i, x\rangle}{\gamma}\right) \right) - \left(m - \frac{1}{2}\right) \\
&= S_{s,r}\left(\frac{\langle v_j, x\rangle}{\gamma}\right) + \left(\sum_{\substack{i \, : \, 1 \le i \le m \\ i \neq j}} S_{s,r}\left(\frac{\langle v_i, x\rangle}{\gamma}\right) \right) - \left(m - \frac{1}{2}\right) \\
&\le -1 + (m-1)\left(1 + \frac{1}{2m}\right) - \left(m - \frac{1}{2}\right) \\
&\le -1/2 \, .
\end{aligned}
$$

□

To prove parts 1 and 2 of Theorem 8 first note that part 4 of Lemma 17 implies that for any x such that $\|x\| \le 1$, $B_{s,r}\left(\frac{\langle v_i,x\rangle}{\gamma}\right)$ is positive. Thus $p(x)$ and $Q(x)$ have the same sign on the unit ball. Consider any x in either $\bigcap_{i=1}^m \left\{x \in \mathbb{R}^d \, : \, \|x\| \le 1, \, \langle v_i, x \rangle \ge \gamma \right\}$ or in $\bigcup_{i=1}^m \left\{x \in \mathbb{R}^d \, : \, \|x\| \le 1, \, \langle v_i, x \rangle \le -\gamma \right\}$. Lemma 18 states that $|Q(x)| \ge 1/2$ and the sign depends on which of the two sets x lies in. Since signs of $Q(x)$ and $p(x)$ are the same, it remains to show that $|p(x)| \ge \frac{1}{4} \cdot 2^{s(s+1)rm}$. Indeed,

$$
\begin{aligned}
|p(x)| &= 2^{-s(s+1)rm+1} \cdot |Q(x)| \prod_{j=1}^m B_{s,r}\left(\frac{\langle v_j, x\rangle}{\gamma}\right) \\
&\ge 2^{-s(s+1)rm+1} \cdot |Q(x)| \left(2^{s(s+1)r}\left(1 - \frac{1}{4m+1}\right)\right)^m \\
&\ge |Q(x)| \ge \frac{1}{2} \quad \text{(Lemma 18)} \, .
\end{aligned}
$$

where we used that $\left(1 - \frac{1}{4m+1}\right)^m \ge e^{-\frac{1}{4}} \ge 1/2$.

To prove part 3 of Theorem 8 note that $\deg(P_s) = 2s+1$. Thus, $\deg(A_{s,r})$ and $\deg(B_{s,r})$ are at most $(2s+1)r$. Therefore, $\deg(p) \le (2s+1)rm$.

It remains to prove part 4 of Theorem 8. For any $i \in \{0, 1, 2, \dots, s\}$ and any $v \in \mathbb{R}^d$ such that $\|v\| \le 1$ define multivariate polynomials

$$
\begin{aligned}
f_{i,v}(x) &= \frac{\langle v, x\rangle}{\gamma} - 2^i \ , \\
q_v(x) &= P_s\left(\frac{\langle v, x\rangle}{\gamma}\right) \ , \\
a_v(x) &= A_{s,r}\left(\frac{\langle v, x\rangle}{\gamma}\right) \ , \\
b_v(x) &= B_{s,r}\left(\frac{\langle v, x\rangle}{\gamma}\right) \ .
\end{aligned}
$$

Note that

$$
q(x) = \left[\sum_{i=1}^m a_{v_i}(x) \prod_{\substack{j \,:\, 1 \le j \le m \\ j \neq i}} b_{v_j}(x)\right] - \left(m - \frac{1}{2}\right) \prod_{j=1}^n b_{v_j}(x) \ .
$$

We bound the norms of these polynomials. We have

$$
\left\|f_{i,v}\right\|^2 = \|v\|^2 / \gamma^2 + 2^{2i} \le 2 \cdot 2^{2s} \ .
$$

where we used that $1/\gamma \le 2^s$ and $\|v\| \le 1$. Since $q_v(x) = f_{i,v}(\frac{\langle v,x\rangle}{\gamma}) \prod_{i=1}^s \left(f_{i,v}(\frac{\langle v,x\rangle}{\gamma})\right)^2$, using part 1 of Lemma 14 we upper bound the norm of q_v as

$$
\begin{aligned}
\|q_v\|^2 &\le (2s+1)^{2s+1} \left\|f_{0,v}\right\|^2 \prod_{i=1}^s \left\|f_{i,v}\right\|^4 \\
&\le (2s+1)^{2s+1} (2 \cdot 2^{2s})^{2s+1} \ .
\end{aligned}
$$

Using parts 3 and 2 of Lemma 14 we upper bound the norm of a_v as

$$
\begin{aligned}
\|a_v\|^2 &\le 2\left\|(q_v)^r\right\|^2 + 2\left\|(q_{-v})^r\right\|^2 \\
&\le 2r^{r(2s+1)} (\|q_v\|^2)^r + 2r^{r(2s+1)} (\|q_{-v}\|^2)^r \\
&\le 4r^{r(2s+1)} \left((2s+1)^{2s+1} (2 \cdot 2^s)^{2s+1}\right)^r \\
&= 4\left(2^{2s} r(4s+2)\right)^{(2s+1)r} \ .
\end{aligned}
$$

The same upper bound holds for $\|b_v\|^2$. Therefore,

$$
\begin{aligned}
\|q\| &\le \left[\sum_{i=1}^m \left\| a_{v_i} \prod_{\substack{j \,:\, 1 \le j \le m \\ j \neq i}} b_{v_j} \right\|\right] + \left(m - \frac{1}{2}\right) \left\| \prod_{j=1}^m b_{v_j} \right\| \\
&\le \left[\sum_{i=1}^m m^{(s+1/2)rm} \|a_{v_i}\| \prod_{\substack{j \,:\, 1 \le j \le m \\ j \neq i}} \|b_{v_j}\|\right] \\
&\qquad + \left(m - \frac{1}{2}\right) m^{(s+1/2)rm} \prod_{j=1}^m \|b_{v_j}\| \\
&\le (2m - 1/2) m^{(s+1/2)rm} \left(4\left(2^{2s} r(4s+2)\right)^{(2s+1)r}\right)^{m/2} \\
&= (2m - 1/2) 2^m \cdot \left(2^{2s} rm(4s+2)\right)^{(s+1/2)rm} \ .
\end{aligned}
$$

Finally, $\|p\| = 2^{-s(s+1)rm+1}\|q\| \le (4m-1)2^m \cdot \left(2^s rm(4s+2)\right)^{(s+1/2)rm}$. The theorem follows.

E. Proof of Theorem 5

Proof of Theorem 5. Since the examples (x_1, y_1), (x_2, y_2), $\ldots$, (x_T, y_T) are weakly linearly separable with margin γ,, there are vectors $w_1, w_2, \ldots, w_K$ satisfying (1) and (2).

Fix any $i \in \{1, 2, \ldots, K\}$. Consider the $K-1$ vectors $(w_i - w_j)/2$ for $j \in \{1, 2, \ldots, K\} \setminus \{i\}$. Note that the vectors have norm at most 1. We consider two cases regarding the relationship between γ_1 and γ_2.

Case 1: $\gamma_1 \ge \gamma_2$**.** In this case, Theorem 7 implies that there exist a multivariate polynomial $p_i : \mathbb{R}^d \to \mathbb{R}$,

$$\deg(p_i) = \lceil \log_2(2K-2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil ,$$

such that all examples x in R_i^+ (resp. R_i^-) satisfy $p_i(x) \ge 1/2$ (resp. $p_i(x) \le -1/2$). Therefore, for all $t = 1, 2, \ldots, T$, if $y_t = i$ then $p_i(x_t) \ge 1/2$, and if $y_t \ne i$ then $p_i(x_t) \le -1/2$, and

$$\|p_i\| \le \left(188 \lceil \log_2(2K-2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right)^{\frac{1}{2}\lceil \log_2(2K-2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil} .$$

By Lemma 9, there exists $c_i \in \ell_2$ such that $\langle c_i, \phi(x) \rangle = p_i(x)$, and

$$\|c_i\|_{\ell_2} \le \left(376 \lceil \log_2(2K-2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right)^{\frac{1}{2}\lceil \log_2(2K-2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil} .$$

Define vectors $u_i \in \ell_2$ as

$$u_i = \frac{1}{\sqrt{K}} \cdot \frac{c_i}{\left(376 \lceil \log_2(2K-2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right)^{\frac{1}{2}\lceil \log_2(2K-2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil}} .$$

Then, $\|u_1\|^2 + \|u_2\|^2 + \cdots + \|u_K\|^2 \le 1$. Furthermore, for all $t = 1, 2, \ldots, T$, $\langle u_{y_t}, \phi(x_t) \rangle \ge \gamma_1$ and for all $j \in \{1, 2, \ldots, K\} \setminus \{y_t\}$, $\langle u_j, \phi(x_t) \rangle \le -\gamma_1$. In other words, $(\phi(x_1), y_1), (\phi(x_2), y_2), \ldots, (\phi(x_T), y_T)$ are strongly linearly separable with margin $\gamma_1 = \max\{\gamma_1, \gamma_2\}$.

Case 2: $\gamma_1 < \gamma_2$**.** In this case, Theorem 8 implies that there exist a multivariate polynomial $q_i : \mathbb{R}^d \to \mathbb{R}$,

$$\deg(q_i) = (2s+1)r(K-1) ,$$

such that all examples x in R_i^+ (resp. R_i^-) satisfy $q_i(x) \ge 1/2$ (resp. $q_i(x) \le -1/2$), and

$$\|q_i\| \le (4K-5)2^{K-1} \cdot \left(2^s r(K-1)(4s+2)\right)^{(s+1/2)r(K-1)} .$$

Recall that here,

$$r = 2\left\lceil \frac{1}{4} \log_2(4K-3) \right\rceil + 1 \quad \text{and} \quad s = \lceil \log_2(1/\gamma) \rceil .$$

Therefore, for all $t = 1, 2, \ldots, T$, if $y_t = i$ then $q_i(x_t) \ge 1/2$, and if $y_t \ne i$ then $q_i(x_t) \le -1/2$.

By Lemma 9, there exists $c_i' \in \ell_2$ such that $\langle c_i', \phi(x) \rangle = p_i(x)$, and

$$\|c_i'\|_{\ell_2} \le (4K-5)2^{K-1} \cdot \left(2^{s+1} r(K-1)(4s+2)\right)^{(s+1/2)r(K-1)} .$$

Define vectors $u'_i \in \ell_2$ as

$$u'_i = \frac{c'_i \cdot \left(2^{s+1} r(K-1)(4s+2)\right)^{-(s+1/2)r(K-1)}}{\sqrt{K}(4K-5)2^{K-1}} \,.$$

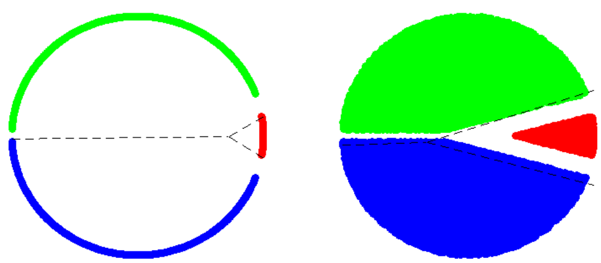
Then, $\left\|u'_1\right\|^2 + \left\|u'_2\right\|^2 + \cdots + \left\|u'_K\right\|^2 \leq 1$. Furthermore, for all $t = 1, 2, \ldots, T$, $\left\langle u'_{y_t}, \phi(x_t) \right\rangle \geq \gamma_2$ and for all $j \in \{1, 2, \ldots, K\} \setminus \{y_t\}$, $\left\langle u'_j, \phi(x_t) \right\rangle \leq -\gamma_2$. In other words, $(\phi(x_1), y_1), (\phi(x_2), y_2), \ldots, (\phi(x_T), y_T)$ are strongly linearly separable with margin $\gamma_2 = \max\{\gamma_1, \gamma_2\}$.

In summary, the examples are strongly linearly separable with margin $\gamma' = \max\{\gamma_1, \gamma_2\}$. Finally, observe that for any $t = 1, 2, \ldots, T$,

$$k(x_t, x_t) = \frac{1}{1 - \frac{1}{2}\|x_t\|^2} \leq 2 \,. \qquad \square$$

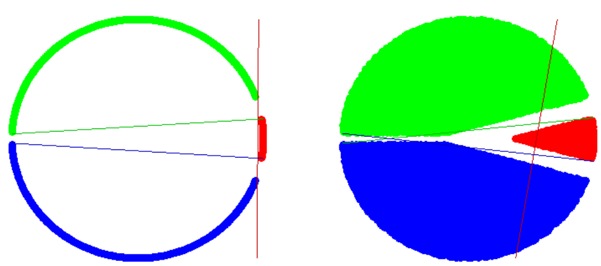
F. Supplementary Materials for Section 6

Figures 6, 7, and 8 show the final decision boundaries learned by each algorithm on the two datasets (Figures 4 and 5), after $T = 5 \times 10^6$ rounds. We used the version of Banditron with exploration rate of 0.02, which explores the most.



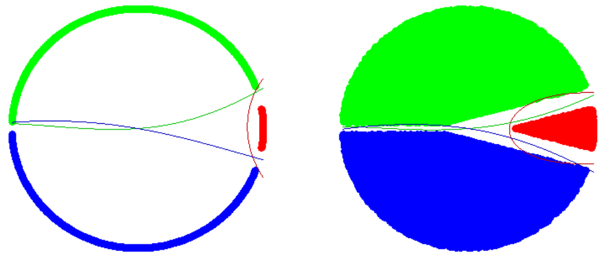
(a) Strongly separable case (b) Weakly separable case

Figure 6. BANDITRON's final decision boundaries



(a) Strongly separable case (b) Weakly separable case

Figure 7. Algorithm 1's final decision boundaries



(a) Strongly separable case (b) Weakly separable case

Figure 8. Algorithm 2 (with rational kernel)'s final decision boundaries

G. Nearest neighbor algorithm

Algorithm 4 NEAREST-NEIGHBOR ALGORITHM

Require: Number of classes K, number of rounds T.
Require: Inner product space $(V, \langle \cdot, \cdot \rangle)$.
Initialize $S \leftarrow \emptyset$
for $t = 1, 2, \ldots, T$: **do**
  if $\min_{(x,y) \in S} \|x_t - x\| \leq \gamma$ **then**
    Find nearest neighbor $(\widetilde{x}, \widetilde{y}) = \text{argmin}_{(x,y) \in S} \|x_t - x\|$
    Predict $\widehat{y}_t = \widetilde{y}$
  else
    Predict $\widehat{y}_t \sim \text{Uniform}(\{1, 2, \ldots, K\})$
    Receive feedback $z_t = \mathbb{1}\left[\widehat{y}_t \neq y_t\right]$
    if $z_t = 0$ **then**
      $S \leftarrow S \cup \left\{(x_t, \widehat{y}_t)\right\}$

In this section we analyze NEAREST-NEIGHBOR ALGORITHM shown as Algorithm 4. The algorithm is based on the obvious idea that, under the weak linear separability assumption, two examples that are close to each other must have the same label. The lemma below formalizes this intuition.

Lemma 19 (Non-separation lemma)**.** *Let $(V, \langle \cdot, \cdot \rangle)$ be a vector space, K be a positive integer and let γ be a positive real number. Suppose $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T) \in V \times \{1, 2, \ldots, K\}$ are labeled examples that are weakly linearly separable with margin γ. For i, j in $\{1, 2, \ldots, T\}$, if $\left\|x_i - x_j\right\|_2 \leq \gamma$ then $y_i = y_j$.*

Proof. Suppose for the sake on contradiction that $y_i \neq y_j$. By Definition 1, there exists vectors $w_1, \ldots, w_K$ such that conditions (1) and (2) are satisfied.

Specifically,

$$\left\langle w_{y_i} - w_{y_j}, x_i \right\rangle \geq \gamma \;,$$
$$\left\langle w_{y_j} - w_{y_i}, x_j \right\rangle \geq \gamma \;.$$

This implies that

$$\left\langle w_{y_i} - w_{y_j}, x_i - x_j \right\rangle \geq 2\gamma \;.$$

On the other hand,

$$\left\langle w_{y_i} - w_{y_j}, x_i - x_j \right\rangle \leq \left\|w_{y_i} - w_{y_j}\right\| \cdot \left\|x_i - x_j\right\| \leq \sqrt{2}\gamma$$

where the first inequality is from Cauchy-Schwartz inequality, the second inequality is from that $\left\|w_{y_i} - w_{y_j}\right\| \leq \sqrt{2(\left\|w_{y_i}\right\|^2 + \left\|w_{y_j}\right\|^2)} \leq \sqrt{2}$ and our assumption on x_i and x_j. Therefore, we reach a contradiction. $\square$

We also need to define several notions. A subset $S \subseteq \mathbb{R}^d$ is called a γ-packing if for any $x, x' \in S$ such that $x \neq x'$ we have $\|x - x'\| > \gamma$. The following lemma is standard. Also recall that $\mathrm{B}(x, R) = \{x' \in \mathbb{R}^d \,:\, \|x' - x\| \le R\}$ denotes the closed ball of radius R centered a point x. For set $S \subseteq \mathbb{R}^d$, denote by $\mathrm{Vol}(S)$ the volume of S.

Lemma 20 (Size of γ-packing)**.** *Let γ and R be positive real numbers. If $S \subseteq \mathrm{B}(\mathbf{0}, R) \subseteq \mathbb{R}^d$ is a γ-packing then*

$$|S| \le \left(\frac{2R}{\gamma} + 1\right)^d .$$

Proof. If S is a γ-packing then $\{\mathrm{B}(x, \gamma/2) \,:\, x \in S\}$ is a collection of disjoint balls of radius γ that fit into $\mathrm{B}(\mathbf{0}, R + \gamma/2)$. Thus,

$$|S| \cdot \mathrm{Vol}(\mathrm{B}(\mathbf{0}, \gamma/2)) \le \mathrm{Vol}(\mathrm{B}(\mathbf{0}, R + \gamma/2))$$

Hence,

$$|S| \le \frac{\mathrm{Vol}(\mathrm{B}(\mathbf{0}, R + \gamma/2))}{\mathrm{Vol}(\mathrm{B}(\mathbf{0}, \gamma/2))} = \left(\frac{R + \gamma/2}{\gamma/2}\right)^d = \left(\frac{2R}{\gamma} + 1\right)^d .$$

□

Theorem 21 (Mistake upper bound for NEAREST-NEIGHBOR ALGORITHM)**.** *Let K and d be positive integers and let γ, R be a positive real numbers. Suppose $(x_1, y_1), \ldots, (x_T, y_T) \in \mathbb{R}^d \times \{1, 2, \ldots, K\}$ are labeled examples that are weakly linearly separable with margin γ and satisfy $\|x_1\|, \|x_2\|, \ldots, \|x_T\| \le R$. Then, the expected number of mistakes made by Algorithm 4 is at most*

$$(K - 1)\left(\frac{2R}{\gamma} + 1\right)^d .$$

Proof. Let M be the number of mistakes made by the algorithm. Let b_t be the indicator that line 7 is executed at time step t, i.e. we fall into the "else" case. Note that if $b_t = 0$, then by Lemma 19, the prediction $\widehat{y}_t$ must equal y_t, i.e. $z_t = 0$. Therefore, $M = \sum_{t=1}^T z_t = \sum_{t=1}^T b_t z_t$. Let $U = \sum_{t=1}^T b_t (1 - z_t)$. Clearly, $|S| = U$. Since $S \subseteq \mathrm{B}(\mathbf{0}, R)$ is a γ-packing, $U = |S| \le (\frac{2R}{\gamma} + 1)^d$.

Note that when $b_t = 1$, $\widehat{y}_t$ is chosen uniformly at random, we have

$$\mathbf{E}[z_t \mid b_t = 1] = \frac{K - 1}{K} .$$

Therefore,

$$\mathbf{E}[M] = \mathbf{E}\left[\sum_{t=1}^T b_t z_t\right] = \frac{K - 1}{K} \, \mathbf{E}\left[\sum_{t=1}^T b_t\right] .$$

On the other hand,

$$\mathbf{E}[U] = \mathbf{E}\left[\sum_{t=1}^T b_t (1 - z_t)\right] = \frac{1}{K} \, \mathbf{E}\left[\sum_{t=1}^T b_t\right] .$$

Therefore,

$$\mathbf{E}[M] = (K - 1)\, \mathbf{E}[U] \le (K - 1)\left(\frac{2R}{\gamma} + 1\right)^d .$$

□

H. NP-hardness of the weak labeling problem

Any algorithm for the bandit setting collects information in the form of so called *strongly labeled* and *weakly labeled* examples. Strongly-labeled examples are those for which we know the class label. Weakly labeled example is an example for which we know that class label can be anything except for a particular one class.

A natural strategy for each round is to find vectors $w_1, w_2, \ldots, w_K$ that linearly separate the examples seen in the previous rounds and use the vectors to predict the label in the next round. More precisely, we want to find both the vectors $w_1, w_2, \ldots, w_K$ and label for each example consistent with its weak and/or strong labels such that $w_1, w_2, \ldots, w_K$ linearly separate the labeled examples. We show this problem is NP-hard even for $K = 3$.

Clearly, the problem is at least as hard as the decision version of the problem where the goal is to determine if such vectors and labeling exist. We show that this problem is NP-complete.

We use symbols $[K] = \{1, 2, \ldots, K\}$ for strong labels and $[\overline{K}] = \{\overline{1}, \overline{2}, \ldots, \overline{K}\}$ for weak labels. Formally, the weak labeling problem can be described as below:

Weak Labeling

Given: Feature-label pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ in $\{0,1\}^d \times \{1, 2, \ldots, K, \overline{1}, \overline{2}, \ldots, \overline{K}\}$.
Question: Do there exist $w_1, w_2, \ldots, w_K \in \mathbb{R}^d$ such that for all $t = 1, 2, \ldots, T$,

$$y_t \in [K] \Longrightarrow \forall i \in [K] \setminus \{y_t\} \quad \langle w_{y_t}, x_t \rangle > \langle w_i, x_t \rangle \ ,$$

and

$$y_t \in [\overline{K}] \Longrightarrow \exists i \in [K] \quad \langle w_i, x_t \rangle > \langle w_{\overline{y_t}}, x_t \rangle \ ?$$

The hardness proof is based on a reduction from the set splitting problem, which is proven to be NP-complete by Lovász (Garey & Johnson, 1979), to our weak labeling problem. The reduction is adapted from (Blum & Rivest, 1993).

Set Splitting

Given: A finite set S and a collection C of subsets c_i of S.
Question: Do there exist disjoint sets S_1 and S_2 such that $S_1 \cup S_2 = S$ and $\forall i, c_i \not\subseteq S_1$ and $c_i \not\subseteq S_2$?

Below we show the reduction. Suppose we are given an instance of the set splitting problem

$$S = \{1, 2, \ldots, N\} \ , C = \{c_1, c_2, \ldots, c_M\} \ .$$

We create the weak labeling instance as follows. Let $d = N + 1$ and $K = 3$. Define $\mathbf{0}$ as the zero vector $(0, \ldots, 0) \in \mathbb{R}^N$ and $\mathbf{e}_i$ as the i-th standard vector $(0, \ldots, 1, \ldots, 0) \in \mathbb{R}^N$. Then we include all the following feature-label pairs:

- Type 1: $(x, y) = ((\mathbf{0}, 1), 3)$,
- Type 2: $(x, y) = ((\mathbf{e}_i, 1), \overline{3})$ for all $i \in \{1, 2, \ldots, N\}$,
- Type 3: $(x, y) = \left(\left(\sum_{i \in c_j} \mathbf{e}_i, 1\right), 3\right)$ for all $j \in \{1, 2, \ldots, M\}$.

For example, if we have $S = \{1, 2, 3\}$, $C = \{c_1, c_2\}$, $c_1 = \{1, 2\}$, $c_2 = \{2, 3\}$, then we create the weak labeling sample set as:

$$\{((0,0,0,1),3), ((1,0,0,1),\overline{3}), ((0,1,0,1),\overline{3}), ((0,0,1,1),\overline{3}), ((1,1,0,1),3), ((0,1,1,1),3)\} \ .$$

The following lemma shows that answering this weak labeling problem is equivalent to answering the original set splitting problem.

Lemma 22. *Any instance of the set splitting problem is a YES instance if and only if the corresponding instance of the weak labeling problem (as described above) is a YES instance.*

Proof. ($\Longrightarrow$) Let S_1, S_2 be the solution of the set splitting problem. Define

$$w_1 = \left(a_1, a_2, \cdots, a_N, -\frac{1}{2}\right),$$

where for all $i \in \{1, 2, \ldots, N\}$, $a_i = 1$ if $i \in S_1$ and $a_i = -N$ if $i \notin S_1$. Similarly, define

$$w_2 = \left(b_1, b_2, \cdots, b_N, -\frac{1}{2}\right),$$

where for all $i \in \{1, 2, \ldots, N\}$, $b_i = 1$ if $i \in S_2$ and $b_i = -N$ if $i \notin S_2$. Finally, define

$$w_3 = (0, 0, \cdots, 0),$$

the zero vector. To see this is a solution for the weak labeling problem, we verify separately for Type 1-3 samples defined above. For Type 1 sample, we have

$$\langle w_3, x\rangle = 0 > -\frac{1}{2} = \langle w_1, x\rangle = \langle w_2, x\rangle .$$

For a Type 2 sample that corresponds to index i, we have either $i \in S_1$ or $i \in S_2$ because $S_1 \cup S_2 = \{1, 2, \ldots, N\}$ is guaranteed. Thus, either $a_i = 1$ or $b_i = 1$. If $a_i = 1$ is the case, then

$$\langle w_1, x\rangle = a_i - \frac{1}{2} = \frac{1}{2} > 0 = \langle w_3, x\rangle ;$$

similarly if $b_i = 1$, we have $\langle w_2, x\rangle > \langle w_3, x\rangle$.
For a Type 3 sample that corresponds to index j, Since $c_j \not\subset S_1$, there exists some $i' \in c_j$ and $i' \notin S_1$. Thus we have $x_{i'} = 1$, $a_{i'} = -N$, and therefore

$$\begin{aligned} \langle w_1, x\rangle &= a_{i'}x_{i'} + \sum_{i \in \{1,2,\ldots,N\}\setminus\{i'\}} a_i x_i - \frac{1}{2} \\ &\le -N + (N-1) - \frac{1}{2} < 0 = \langle w_3, x\rangle \ . \end{aligned}$$

Because $c_j \not\subset S_2$ also holds, we also have $\langle w_2, x\rangle < \langle w_3, x\rangle$. This direction is therefore proved.

($\Longleftarrow$) Given the solution w_1, w_2, w_3 of the weak labeling problem, we define

$$\begin{aligned} S_1 &= \Big\{ i \in \{1, 2, \ldots, n\} \ : \ \big\langle w_1 - w_3, (\mathbf{e}_i, 1)\big\rangle > 0 \Big\}, \\ S_2 &= \Big\{ i \in \{1, 2, \ldots, n\} \ : \ \big\langle w_2 - w_3, (\mathbf{e}_i, 1)\big\rangle > 0 \text{ and } i \notin S_1 \Big\}. \end{aligned}$$

It is not hard to see $S_1 \cap S_2 = \emptyset$ and $S_1 \cup S_2 = \{1, 2, \ldots, N\}$. The former is because S_2 only includes elements that are not in S_1. For the latter, note that $(\mathbf{e}_i, 1)$ is the feature vector for Type 2 samples. Because Type 2 samples all have label $\overline{3}$, for any $i \in \{1, 2, \ldots, N\}$, one of the following must hold: $\big\langle w_1 - w_3, (\mathbf{e}_i, 1)\big\rangle > 0$ or $\big\langle w_2 - w_3, (\mathbf{e}_i, 1)\big\rangle > 0$. This implies $i \in S_1$ or $i \in S_2$.

Now we show $\forall j$, $c_j \not\subset S_1$ and $c_j \not\subset S_2$ by contradiction. Assume there exists some j such that $c_j \subset S_1$. By our definition of S_1, we have $\big\langle w_1 - w_3, (\mathbf{e}_i, 1)\big\rangle > 0$ for all $i \in c_j$. Therefore,

$$\sum_{i \in c_j} \big\langle w_1 - w_3, (\mathbf{e}_i, 1)\big\rangle = \left\langle w_1 - w_3, \left(\sum_{i \in c_j} \mathbf{e}_i, |c_j|\right)\right\rangle > 0.$$

Because Type 1 sample has label 3, we also have

$$\big\langle w_1 - w_3, (\mathbf{0}, 1) \big\rangle < 0.$$

Combining the above two inequalities, we get

$$\left\langle w_1 - w_3, \left(\sum_{i \in c_j} \mathbf{e}_i, 1 \right) \right\rangle = \left\langle w_1 - w_3, \left(\sum_{i \in c_j} \mathbf{e}_i, |c_j| \right) \right\rangle - (|c_j| - 1) \big\langle w_1 - w_3, (\mathbf{0}, 1) \big\rangle > 0 \, .$$

Note that $\left(\sum_{i \in c_j} \mathbf{e}_i, 1\right)$ is a feature vector for Type 3 samples. Thus the above inequality contradicts that Type 3 samples have label 3. Therefore, $c_j \not\subset S_1$. If we assume there exists some $c_j \subset S_2$, same arguments apply and also lead to contradiction. □

I. Mistake lower bound for ignorant algorithms

In this section, we prove a mistake lower bound for a family of algorithms called *ignorant algorithms*. Ignorant algorithms ignore the examples on which they make mistakes. This assumption seems strong, but as we will explain below, it is actually natural, and several recently proposed bandit linear classification algorithms that achieve $\sqrt{T}$ regret bounds belong to this family, e.g., SOBA (Beygelzimer et al., 2017), OBAMA (Foster et al., 2018). Also, NEAREST-NEIGHBOR ALGORITHM (Algorithm 4) presented in Appendix G is an ignorant algorithm.

Under the assumption that the examples lie in in the unit ball of $\mathbb{R}^d$ and are weakly linearly separable with margin γ, we show that any ignorant algorithm must make at least $\Omega\left(\left(\frac{1}{160\gamma}\right)^{(d-2)/4}\right)$ mistakes in the worst case. In other words, an algorithm that achieves a better mistake bound cannot ignore examples on which it makes a mistake and it must make a meaningful update on such examples.

To formally define ignorant algorithms, we define the conditional distribution from which an algorithm draws its predictions. Formally, given an algorithm $\mathcal{A}$ and an adversarial strategy, we define

$$p_t(y|x) = \Pr[y_t = y \mid (x_1, y_1), (x_2, y_2) \dots, (x_{t-1}, y_{t-1}), x_t = x] \, .$$

In other words, in any round t, conditioned on the past $t-1$ rounds, the algorithm $\mathcal{A}$ chooses y_t from probability distribution $p_t(\cdot|x_t)$. Formally, p_t is a function $p : \{1, 2, \dots, K\} \times \mathbb{R}^d \to [0, 1]$ such that $\sum_{y=1}^{K} p_t(y|x) = 1$ for any $x \in \mathbb{R}^d$.

Definition 23 (Ignorant algorithm)**.** *An algorithm $\mathcal{A}$ for* ONLINE MULTICLASS LINEAR CLASSIFICATION WITH BANDIT FEEDBACK *is called* ignorant *if for every $t = 1, 2, \dots, T$, p_t is determined solely by the sequence (x_{a_1}, y_{a_1}),(x_{a_2}, y_{a_2}), $\dots$, (x_{a_n}, y_{a_n}) of labeled examples from the rounds $1 \le a_1 < a_2 < \cdots < a_n < t$ in which the algorithm makes a correct prediction.*

An equivalent definition of an ignorant algorithm is that the memory state of the algorithm does not change after it makes a mistake. Equivalently, the memory state of an ignorant algorithm is completely determined by the sequence of labeled examples on which it made correct prediction.

To explain the definition, consider an ignorant algorithm $\mathcal{A}$. Suppose that on a sequence of examples (x_1, y_1), (x_2, y_2), $\dots$, (x_{t-1}, y_{t-1}) generated by some adversary the algorithm $\mathcal{A}$ makes correct predictions in rounds $a_1, a_2, \dots, a_n$ where $1 \le a_1 < a_2 < \cdots < a_n < t$ and errors on rounds $\{1, 2, \dots, t-1\} \setminus \{a_1, a_2, \dots, a_n\}$. Suppose that on another sequence of examples $(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_{s-1}, y'_{s-1})$ generated by another adversary the algorithm $\mathcal{A}$ makes correct predictions in rounds $b_1, b_2, \dots, b_n$ where $1 \le b_1 < b_2 < \cdots < b_n < s$ and errors on rounds $\{1, 2, \dots, s-1\} \setminus \{b_1, b_2, \dots, b_n\}$. Futhermore, suppose

$$\begin{gathered}
(x_{a_1}, y_{a_1}) = (x'_{b_1}, y'_{b_1}) \, , \\
(x_{a_2}, y_{a_2}) = (x'_{b_2}, y'_{b_2}) \, , \\
\vdots \\
(x_{a_n}, y_{a_n}) = (x'_{b_2}, y'_{b_n}) \, .
\end{gathered}$$

Then, as $\mathcal{A}$ is ignorant,

$$\Pr[y_t = y \mid (x_1, y_1), (x_2, y_2) \ldots, (x_{t-1}, y_{t-1}), x_t = x] = \Pr[y'_t = y \mid (x'_1, y'_1), (x'_2, y'_2) \ldots, (x'_{t-1}, y'_{t-1}), x'_t = x].$$

Note that the sequences (x_1, y_1), (x_2, y_2), $\ldots$, (x_{t-1}, y_{t-1}) and (x'_1, y'_1), (x'_2, y'_2), $\ldots$, (x'_{s-1}, y'_{s-1}) might have different lengths and and $\mathcal{A}$ might error in different sets of rounds. As a special case, if an ignorant algorithm makes a mistake in round t then $p_{t+1} = p_t$.

Our main result is the following lower bound on the expected number of mistakes for ignorant algorithms.

Theorem 24 (Mistake lower bound for ignorant algorithms)**.** *Let $\gamma \in (0, 1)$ and let d be a positive integer. Suppose $\mathcal{A}$ is an ignorant algorithm for* ONLINE MULTICLASS LINEAR CLASSIFICATION WITH BANDIT FEEDBACK. *There exists T and an adversary that sequentially chooses labeled examples $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T) \in \mathbb{R}^d \times \{1, 2\}$ such that the examples are strongly linearly separable with magin γ and $\|x_1\|, \|x_2\|, \ldots, \|x_T\| \le 1$, and the expected number of mistakes made by $\mathcal{A}$ is at least*

$$\frac{1}{10}\left(\frac{1}{160\gamma}\right)^{\frac{d-2}{4}} .$$

Before proving the theorem, we need the following lemma.

Lemma 25. *Let $\gamma \in (0, \frac{1}{160})$, let d be a positive integer and let $N = (\frac{1}{2\sqrt{40\gamma}})^{d-2}$. There exist vectors $u_1, u_2, \ldots, u_N$, $v_1, v_2, \ldots, v_N$ in $\mathbb{R}^d$ such that for all $i, j \in \{1, 2, \ldots, N\}$,*

$$\begin{aligned} \|u_i\| &\le 1 \ , \\ \|v_j\| &\le 1 \ , \\ \langle u_i, v_j \rangle &\ge \gamma, \quad \text{if } i = j, \\ \langle u_i, v_j \rangle &\le -\gamma, \quad \text{if } i \neq j. \end{aligned}$$

Proof. By Lemma 6 of Long (1995), there exists vectors $z_1, z_2, \ldots, z_N \in \mathbb{R}^{d-1}$ such that $\|z_1\| = \|z_2\| = \cdots = \|z_N\| = 1$ and the angle between the vectors is $\measuredangle(z_i, z_j) \ge \sqrt{40\gamma}$ for $i \neq j$, $i, j \in \{1, 2, \ldots, N\}$. Since $\cos\theta \le 1 - \theta^2/5$ for any $\theta \in [-\pi, \pi]$, this implies that

$$\begin{aligned} \langle z_i, z_j \rangle &= 1, \quad \text{if } i = j, \\ \langle z_i, z_j \rangle &\le 1 - 8\gamma, \quad \text{if } i \neq j. \end{aligned}$$

Define $v_i = (\frac{1}{2}z_i, \frac{1}{2})$, and $u_i = (\frac{1}{2}z_i, -\frac{1}{2}(1-4\gamma))$ for all $i \in \{1, 2, \ldots, N\}$. It can be easily checked that for all i, $\|v_i\| \le 1$ and $\|u_i\| \le 1$. Additionally,

$$\langle u_i, v_j \rangle = \frac{1}{4}\langle z_i, z_j \rangle - \frac{1-4\gamma}{4} .$$

Thus,

$$\begin{aligned} \langle u_i, v_j \rangle &\ge \gamma, \quad \text{if } i = j, \\ \langle u_i, v_j \rangle &\le -\gamma, \quad \text{if } i \neq j. \end{aligned}$$

$\square$

Proof of Theorem 24. We consider the strategy for the adversary described in Algorithm 5.

Let τ be the time step t in which the adversary sets PHASE $\leftarrow 2$. If the adversary never sets PHASE $\leftarrow 2$, we define $\tau = T + 1$. Then,

$$\mathbf{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\widehat{y}_t \neq y_t\right]\right] \ge \mathbf{E}\left[\sum_{t=1}^{\tau-1} \mathbb{1}\left[\widehat{y}_t \neq y_t\right]\right] + \mathbf{E}\left[\sum_{t=\tau}^{T} \mathbb{1}\left[\widehat{y}_t \neq y_t\right]\right] .$$

We upper bound each of last two terms separately.

Algorithm 5 ADVERSARY'S STRATEGY

Define $T = N$ and $v_1, v_2, \ldots, v_N$ as in Lemma 25.
Define $q_0 = \frac{1}{\sqrt{T}}$.
Initialize PHASE $= 1$.
for $t = 1, 2, \ldots, T$ **do**
  if PHASE $= 1$ **then**
    if $p_t(1|v_t) < 1 - q_0$ **then**
      $(x_t, y_t) \leftarrow (v_t, 1)$
    else
      $(x_t, y_t) \leftarrow (v_t, 2)$
      PHASE $\leftarrow 2$
  else
    $(x_t, y_t) \leftarrow (x_{t-1}, y_{t-1})$

In rounds $1, 2, \ldots, \tau - 1$, the algorithm predicts the incorrect class 2 with probability at least q_0. Thus,

$$\mathbf{E}\left[\sum_{t=1}^{\tau-1} \mathbb{1}\left[\widehat{y}_t \neq y_t\right]\right] = q_0\, \mathbf{E}[(\tau - 1)]\ . \tag{18}$$

In rounds $\tau, \tau+1, \ldots, T$, all the examples are the same and are equal to $(v_\tau, 2)$. Let s be the first time step t such that $t \geq \tau$ and the algorithm makes a correct prediction. If the algorithm makes mistakes in all rounds $\tau, \tau+1, \ldots, T$, we define $s = T + 1$. By definition the algorithm makes mistakes in rounds $\tau, \tau+1, \ldots, s-1$. Therefore,

$$\mathbf{E}\left[\sum_{t=\tau}^{T} \mathbb{1}\left[\widehat{y}_t \neq y_t\right]\right] \geq \mathbf{E}[s - \tau]. \tag{19}$$

Since the algorithm is ignorant, conditioned on τ and $q \triangleq p_\tau(2|v_\tau)$, $s - \tau$ follows a truncated geometric distribution with parameter q (i.e., $s - \tau$ is 0 with probability q, 1 with probability $(1-q)q$, 2 with probability $(1-q)^2 q, \ldots$). Its conditional expectation can be calculated as follows:

$$\begin{aligned}
\mathbf{E}[s - \tau \mid \tau, q] &= \sum_{i=1}^{T+1-\tau} i \times \Pr[s - \tau = i \mid \tau, q] \\
&= \sum_{j=1}^{T+1-\tau} \Pr[s - \tau \geq j \mid \tau, q] \\
&= \sum_{j=1}^{T+1-\tau} (1-q)^j \geq \sum_{j=1}^{T+1-\tau} (1-q_0)^j \\
&= \frac{1-q_0}{q_0}\left(1 - (1-q_0)^{T-\tau+1}\right).
\end{aligned}$$

Therefore, by the tower property of conditional expectation,

$$\mathbf{E}[s - \tau \mid \tau] = \mathbf{E}\left[\mathbf{E}\left[s - \tau \mid \tau, q\right] \,\middle|\, \tau\right] \geq \frac{1-q_0}{q_0}\left(1 - (1-q_0)^{T-\tau+1}\right).$$

Combining this fact with Equations (18) and (19), we have that

$$\begin{aligned}
\mathbf{E}\left[\sum_{t=1}^{T} \mathbb{1}\left[\widehat{y}_t \neq y_t\right]\right] &\geq q_0\, \mathbf{E}[\tau - 1] + \mathbf{E}\left[\frac{1-q_0}{q_0}\left(1 - (1-q_0)^{T-\tau+1}\right)\right] \\
&= \mathbf{E}\left[q_0(\tau - 1) + \frac{1-q_0}{q_0}\left(1 - (1-q_0)^{T-\tau+1}\right)\right]\ .
\end{aligned}$$

We lower bound the last expression by considering two cases for τ. If $\tau \geq \frac{1}{2}T + 1$, then the last expression is lower bounded by $\frac{1}{2}q_0 T = \frac{1}{2}\sqrt{T}$. If $\tau < \frac{1}{2}T + 1$, it is lower bounded by

$$\begin{aligned}
&\frac{1-q_0}{q_0}\left(1-(1-q_0)^{\frac{1}{2}T}\right) \\
&= \frac{1-q_0}{q_0}\left(1-(1-q_0)^{\frac{1}{2q_0^2}}\right) \\
&\geq \frac{1-\frac{1}{\sqrt{2}}}{q_0}\left(1-\frac{1}{\sqrt{e}}\right) \\
&\geq \frac{1}{10}\sqrt{T}\ .
\end{aligned}$$

Observe that in phase 1, the labels are equal to 1 and in phase 2 the labels are equal to 2. Note that $(x_\tau, y_\tau) = (x_{\tau+1}, y_{\tau+1}) = \cdots = (x_T, y_T) = (v_\tau, 2)$. Consider the vectors $u_1, u_2, \ldots, u_N$ as defined in Lemma 25. We claim that $w_1 = -u_\tau/2$ and $w_2 = u_\tau/2$ satisfy the conditions of strong linear separability.

Clearly $\|w_1\|^2 + \|w_2\|^2 \leq (\|w_1\| + \|w_2\|)^2 \leq (\frac{1}{2} + \frac{1}{2})^2 \leq 1$. By Lemma 25, we have $\left\langle w_2/2, x_t \right\rangle = \left\langle u_\tau/2, v_\tau \right\rangle \geq \gamma/2, \forall t \geq \tau$ and $\left\langle w_2/2, x_t \right\rangle = \left\langle u_\tau/2, v_t \right\rangle \leq -\gamma/2$ for all $t < \tau$. Similarly, $\left\langle w_1/2, x_t \right\rangle \leq -\gamma/2$ for all $t \geq \tau$ and $\left\langle w_1/2, x_t \right\rangle \geq \gamma/2$ for all $t < \tau$. Thus, the examples are strongly linearly separable with margin γ. $\square$