THESIS

BANDIT MULTICLASS LINEAR CLASSIFICATION FOR THE GROUP LINEAR SEPARABLE CASE

CHAYUTPONG PROMPAK

GRADUATE SCHOOL, KASETSART UNIVERSITY
2017

THESIS

BANDIT MULTICLASS LINEAR CLASSIFICATION FOR THE GROUP LINEAR SEPARABLE CASE

CHAYUTPONG PROMPAK

A Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree of
Master of Engineering (Computer Engineering)
Graduate School, Kasetsart University
2017

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

BANDIT MULTICLASS LINEAR CLASSIFICATION FOR THE GROUP LINEAR SEPARABLE CASE

INTRODUCTION

Classification is the classical problem in machine learning, that is there are inputs and each of them labeled by a class, the task is to receive only inputs and predicts what classes they belong to. There are various types of classifiers, the common one is linear classifiers.

Linear classifiers find a relationship between classes that are represented by linear combination and input, or in the other words, the classifier views an input as a d dimensional vector and separates the vector from the others with $d-1$ dimensional hyperplane. Some of the well-known examples are support-vector machines(SVMs) and perceptron. Everything seems to be fine, however, some of the data can't be separated by the hyperplane but linear classification can also classify non-linear data by using the kernel. The concept is to mapping data into higher-dimensional space which should be made the separable easier.

Generally, classifier accuracy is evaluated by error rate. Although for online learning, the classifier classifies the input and updates itself each time. Mistake bound is an alternative approach that looks appropriate for evaluation. Moreover, there is one more feedback setting for the classification problem. After the classifier predicts the input class then receives feedback only correct if the prediction is correct and wrong if otherwise. In full-information feedback the multiclass perceptron, Crammer and Singer (2003) is $\lfloor 2(R/\gamma)^2 \rfloor$ mistakes and any algorithms must make at least $\frac{1}{2}\lfloor (R/\gamma)^2 \rfloor$ mistakes. In bandit feedback, Beygelzimer *et al.* (2019) provide a simple algorithm and show that the mistake bound is $O(K(R/\gamma)^2)$ in expectation, and for any randomized algorithm must make $\Omega(K(R/\gamma)^2)$ mistakes.

Beygelzimer *et al.* (2019) considered two notions of linear separability, weak and strong linear separability from Crammer and Singer (2003) and employed rational kernel to deal with examples under the weakly linearly separable condition, and obtained the mistake bound of $\min(K \cdot 2^{\tilde{O}(K \log^2(1/\gamma))}, K \cdot 2^{\tilde{O}(\sqrt{1/\gamma} \log K)})$ we refine the notion of weak linear separability to support the notion of class grouping, called group weak linear separable condition. This situation may arise from the fact that class structures contain inherent grouping. We show that under this condition, we can also use the rational ke rnel and obtain the mistake bound of $K \cdot 2^{\tilde{O}(\sqrt{1/\gamma} \log L)}$, where $L \leq K$ represents the number of groups.

LITERATURE REVIEW

1. Online multiclass classification

For classification in online version, an input x_t comes at time t , classifier predicts $\hat{y}_t$ then receives feedback information y_t that is the class of x_t. Let K be the number of classes, T be the number of rounds and $y_t \in \{1, 2, \ldots, K\}$. The protocal(fig 1) does until time T.

begin
  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 $\hat{y}_t \in \{1, 2, \ldots, K\}$.
    Observe feedback y_t.
  end
end

Figure 1 Online Multiclass Classification Protocal (Crammer and Singer (2003))

2. Perceptron

Perceptron (Rosenblatt (1958)) is a binary classifier that answers only input x is in the class or not, A vector w represented as a class and the answer is depends on the inner product of x and w (fig 2). The perceptron update w only when the answer is wrong and makes mistake at most $\left\lfloor \left(\frac{R}{\gamma}\right)^2 \right\rfloor$ (theorem 1).

Theorem 1 (Rosenblatt (1958))**.** *Let $(V, \cdot, \cdot)$ 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_K, y_K)$ is a sequence of labeled examples in $V \times \{+1, -1\}$ and some $\gamma \geq 0$ and $\|x^1\|, \|x^2\|, \ldots, \|x^T\| \leq R$ then PERCEPTRON algorithm makes at most $\left\lfloor \left(\frac{R}{\gamma}\right)^2 \right\rfloor$ mistakes.*

Data: Number of rounds T, $y_t \in \{+1, -1\}$
Data: Inner product space $(V, \langle \cdot, \cdot \rangle)$
begin
  Initialize $w^{(1)} = 0$
  for $t = 1, 2, \ldots, T$ **do**
    Predict class label $\hat{y}_t \begin{cases} +1 & \text{if } \langle x_t, w^{(t)} \rangle \geq 0 \\ -1 & \text{otherwise} \end{cases}$.
    Observe feedback y_t.
    if $\hat{y}_t \neq y_t$ **then**
      Update $w^{(t+1)} = w^{(t)} + \hat{y}_t x_t$
    end
  end
end

Figure 2 Perceptron Algorithm (Rosenblatt (1958))

3. Multiclass Perceptron

Let K be the number of classes, multiclass perceptron construct from K perceptrons and update depends on y_t, see in (fig 3). The multiclass perceptron makes mistake at most $\left\lfloor 2\left(\frac{R}{\gamma}\right)^2 \right\rfloor$ (theorem 2).

Theorem 2 (Crammer and Singer (2003))**.** *Let $(V, \cdot, \cdot)$ 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_K, y_K)$ 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 $\left\lfloor 2\left(\frac{R}{\gamma}\right)^2 \right\rfloor$ mistakes.*

4. Linear seperability

We restate the definitions for strong and weak linear separability by Beygelzimer *et. al.* Beygelzimer *et al.* (2019) here. We use the common notation that $[K] = \{1, 2, \ldots, K\}$.

Data: Number of classes K, number of rounds T
Data: Inner product space $(V, \langle \cdot, \cdot \rangle)$
begin
  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 $\hat{y}_t = \text{argmax}_{i \in \{1,2,\ldots,K\}} \left\langle w_t^{(i)}, x_t \right\rangle$.
    Observe $y_t \in \{1, 2, \ldots, K\}$.
    if $\hat{y}_t \neq y_t$ **then**
      Set $w_i^{(t+1)} = w_i^{(t)}$
      For all $i \in \{1, 2, \ldots, K\} \setminus \{y_t, \hat{y}_t\}$
      Update $w_{y_t}^{(t+1)} = w_{y_t}^{(t)} + x_t$
      Update $w_{\hat{y}_t}^{(t+1)} = w_{\hat{y}_t}^{(t)} - x_t$
    else
      Set $w_i^{(t+1)} = w_i^{(t)}$ for all $i \in \{1, 2, \ldots, K\}$
    end
  end
end

Figure 3 Multiclass Perceptron Algorithm (Crammer and Singer (2003))

4.1 Strongly linear seperable

The examples lie in an inner product space $(V, \langle \cdot, \cdot \rangle)$. Let K be the number of classes and let γ be a positive real number. Labeled examples

$$(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T) \in V \times [K]$$

are *strongly linear separable with margin* γ if there exist vectors $w_1, w_2, \ldots, w_K \in V$ such that for all $t \in [T]$,

$$\langle x_t, w_{y_t} \rangle \geq \gamma/2,$$

and

$$\langle x_t, w_i \rangle \leq -\gamma/2,$$

for $i \in [K] \setminus \{y_t\}$, and $\sum_{i=1}^{K} \|w_i\|^2 \leq 1$.

4.2 Weakly linear seperable

On the other hand, the labeled examples are *weakly linear separable with margin* γ if there exist vectors $w_1, w_2, \ldots, w_K \in V$ such that for all $t \in [T]$,

$$\langle x_t, w_{y_t} \rangle \geq \langle x_t, w_i \rangle + \gamma,$$

for $i \in [K] \setminus \{y_t\}$, and $\sum_{i=1}^{K} \|w_i\|^2 \leq 1$.

5. Kernel

We give an overview of the kernel methods (see Shawe-Taylor and Cristianini (2004) for expositions) and the rational kernel Shalev-Shwartz *et al.* (2011).

The kernel method is a standard approach to extend linear classification algorithms that use only inner products to handle the notions of "distance" between pairs of examples to nonlinear classification. A *positive definite kernel* (or *kernel*) is a function of the form $k : X \times X \to \mathbb{R}$ for some set X such that the matrix $[k(x_i, x_j)]_{i,j=1}^{m}$ is symmetric positive definite for any set of m examples $x_1, x_2, \ldots, x_m \in X$. It is known that for every kernel k, there exists some inner product space $(V, \langle \cdot, \cdot \rangle)$ and a feature map $\phi : X \to V$ such that $k(x, x') = \langle \phi(x), \phi(x') \rangle$. Therefore, a linear learning algorithm can essentially non-linearly map every example into V and work in V instead of the original space without explicitly working with ϕ using k. This can be very helpful when the dimension of V is infinite.

As in Beygelzimer *et al.* Beygelzimer *et al.* (2019), we use the rational kernel. Assume that examples are in $\mathbb{R}^d$. Denote by $\mathrm{B}(0, 1)$ a unit ball centered at 0 in $\mathbb{R}^d$. The *rational kernel* $k : B(0, 1) \times B(0, 1) \to \mathbb{R}$ is defined as

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

Given $x, x' \in \mathbb{R}^d$, $k(x, x')$ can be computed in $O(d)$ time.

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

$$(\phi(x_1, x_2, \ldots, x_d))_{(\alpha_1,\alpha_2,\ldots,\alpha_d)} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d} \cdot \sqrt{2^{-(\alpha_1+\alpha_2+\cdots+\alpha_d)} \binom{\alpha_1+\alpha_2+\cdots+\alpha_d}{\alpha_1, \alpha_2, \ldots, \alpha_d}}, \quad (1)$$

where $\binom{\alpha_1+\alpha_2+\cdots+\alpha_d}{\alpha_1,\alpha_2,\ldots,\alpha_d} = \frac{(\alpha_1+\alpha_2+\cdots+\alpha_d)!}{\alpha_1!\alpha_2!\cdots\alpha_d!}$ is the multinomial coefficient. It can be verified that k is the kernel with its feature map ϕ to ℓ_2 and for any $x \in \mathrm{B}(0,1)$, $\phi(x) \in \ell_2$.

6. Multiclass Linear Classification

Beygelzimer *et al.* Beygelzimer *et al.* (2019) presented a learning algorithm for the strongly linearly separable examples based using K copies of the BINARY PERCEPTRON. They obtained a mistake bound of $O(K(R/\gamma)^2)$ when the examples are from $\mathbb{R}^d$ with maximum norm R with margin γ.

Their approach for dealing the weakly linear separable case is to use the kernel method. They introduced the KERNELIZED BANDIT ALGORITHM (Algorithm 4) and proved the following theorem.

Theorem 3 (Theorem 4 from Beygelzimer *et al.* (2019))**.** *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 $k : X \times X \to \mathbb{R}$, where $k(x, x') = \langle \phi(x), \phi(x') \rangle$, be the kernel. 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,*

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

Figure 4 Kernelized Bandit Algorithm (Beygelzimer *et al.* (2019))

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 the KERNELIZED BANDIT ALGORITHM *makes is at most* $(K-1)\lfloor 4(R/\gamma)^2 \rfloor$.

The key theorem for establishing the mistake bound is the following margin transformation theorem based on the rational kernel.

Theorem 4 (Theorem 5 from Beygelzimer *et al.* (2019))**.** *(Margin transformation from Beygelzimer* et al. *(2019)). Let* $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T) \in \mathrm{B}(0,1) \times [K]$ *be a sequence of labeled examples that is weakly linear separable with margin* $\gamma > 0$*. Let* ϕ *defined as in (1) let*

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

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

where $r = 2\lceil \frac{1}{4}\log_2(4K-3)\rceil + 1$ *and* $s = \lceil \log_2(2/\gamma)\rceil$*. Then the feature map* ϕ *makes the sequence* $(\phi(x_1), y_1), (\phi(x_2), y_2), \ldots, (\phi(x_T), y_T)$ *strongly linearly separable with margin* $\gamma' = \max\{\gamma_1, \gamma_2\}$*. Also for all* t*,* $k(x_t, x_t) \leq 2$*.*

This implies the following mistake bound.

Corollary 1 (Corollary 6 from Beygelzimer *et al.* (2019))**.** *(Mistake upper bound from Beygelzimer* et al. *(2019)). The mistake bound made by Algorithm 4 when the examples are weakly linearly separable with margin* γ *is at most* $\min(2^{\tilde{O}(K \log^2(1/\gamma))}, 2^{\tilde{O}(\sqrt{1/\gamma}\log K)})$.

Beygelzimer *et al.* Beygelzimer *et al.* (2019) gave two margin transformation proofs. In this paper, we only provide one margin transformation based on the Chebyshev polynomials (Theorem 7 from Beygelzimer *et al.* (2019)).

7. Separating polynomials

This section proves Theorem 5. As in Beygelzimer *et al.* (2019) and Klivans and Servedio (2008), we use the Chebyshev polynomials Mason and Handscomb (2002)

$T_n(\cdot)$ defined as follows.

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

The following two lemmas are from Beygelzimer *et al.* (2019).

Lemma 1 (from Lemma 15 in Beygelzimer *et al.* (2019))**.** *(Properties of Chebyshev polynomials) Chebyshev polynomials satisfy*

1. $deg(T_n) = n$ *for all* $n \geq 0$.
2. *If* $n \geq 1$*, 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* $n \geq 0$.
4. $T_n(\cosh(\theta)) = \cosh(n\theta)$ *for all* $\theta \in \mathbb{R}$ *and all* $n \geq 0$.
5. $|T_n(z)| \leq 1$ *for all* $z \in [-1, 1]$ *and all* $n \geq 0$.
6. $T_n(z) \geq 1 + n^2(z-1)$ *for all* $z \geq 1$ *and all* $n \geq 0$.
7. $\|T_n\| \leq (1+\sqrt{2})^n$ *for all* $n \geq 0$.

Lemma 2 (from Lemma 14 in Beygelzimer *et al.* (2019))**.** *(Properties of 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,* $\|p\|^2 \leq n^{\sum_{j=1}^{n} deg(p_j)} \prod_{j=1}^{n} \|p_j\|^2$.
2. *Let* q *be a multivariate polynomial of degree at most* s *and let* $p(x) = (q(x))^n$. *Then,* $\|p\|^2 \leq n^{ns}\|q\|^{2n}$.
3. *Let* $p_1, p_2, \ldots, p_n$ *be multivariate polynomials. Then,* $\left\|\sum_{j=1}^{n} p_j\right\|^2 \leq n \sum_{j=1}^{n} \|p_j\|^2$.

METHODOLOGIES

8. Intra-group boundaries

We first prove a structural property of intra-group classes. The following lemma shows that it is possible to separate one class from the rest in the same group using only lower and upper thresholds. This is independent of the number of classes in that group.

Lemma 3. *For any group $i \in [L]$, for any class $y \in G_i$, there exists reals $b_i \leq t_i$ such that for all $t \in [T]$ such that (1) when $y_t = y$,*

$$b_i + \gamma \leq \langle u'_i, x_t \rangle \leq t_i - \gamma;$$

and (2) when $g(y_t) = g(y)$ but $y_t \neq y$, either

$$\langle x_t, u'_i \rangle \leq b_i - \gamma,$$

or

$$\langle x_t, u'_i \rangle \geq t_i + \gamma.$$

Proof. lemma 3 Let $S_y = \{(x_j, y_j) : y_j = y, 1 \leq j \leq T\}$ be the set of examples with label y. Let $b_i = \min_{(x,y)\in S_y} \langle x, u'_i \rangle - \gamma$ and $t_i = \max_{(x,y)\in S_y} \langle x, u'_i \rangle + \gamma$. The lemma follows from the definition of group weakly linear separability. □

9. Group weakly linear separability

We now define group weakly linear separability. Let $\mathcal{G} = \{G_1, G_2, \ldots, G_L\}$ be a partition of $[K]$, i.e., $G_i \subseteq [K]$ for all i, $G_i \cap G_j = \emptyset$ for $i \neq j$, and $\bigcup G_i = [K]$. Let $g : [K] \to [L]$ be a mapping function such that $g(i) \mapsto j$ iff $i \in G_j$. We say that the

labeled examples

$$(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T) \in V \times [K]$$

are *group weakly linear separable with margin γ under $\mathcal{G}$* if

1. there exist vectors $u_1, u_2, \ldots, u_L \in V$ such that $\sum_{i=1}^{L} \|u_i\|^2 \leq 1$, and, for all $t \in [T]$,

   $$\langle x_t, u_{g(y_t)} \rangle \geq \langle x_t, u_p \rangle + \gamma,$$

   for all $p \in [L] \setminus \{g(y_t)\}$,

2. there exist vectors $u'_1, u'_2, \ldots, u'_L \in V$ such that $\sum_{i=1}^{L} \|u'_i\|^2 \leq 1$, and, for all $t \in [T], t' \in [T]$ such that $y_t \neq y_{t'}$ and $g(y_t) = g(y_{t'})$, either

   $$\langle x_t, u'_{g(y_t)} \rangle \geq \langle x_{t'}, u'_{g(y_t)} \rangle + 2\gamma,$$

   or

   $$\langle x_t, u'_{g(y_t)} \rangle \leq \langle x_{t'}, u'_{g(y_t)} \rangle - 2\gamma.$$

Note that vectors u_i's define inter-group hyperplanes, while each u'_i defines intra-group boundaries. Also note that, to simplify our proofs, the "margin" between intra-group classes is 2γ; this would create the $+\gamma$ and $-\gamma$ gaps that already exist between groups.

To illustrate the idea, Fig. 5 shows 3 sets of examples.

10. Margin transformation

This section is devoted to the proofs of Theorem 6. A key property of the space ℓ_2 is that it contains all multivariate polynomials and the rational kernel k allows us to work in that space. The following lemma is from Beygelzimer *et al.* (2019).

Lemma 4 (from Lemma 9 in Beygelzimer *et al.* (2019))**.** *(Norm bound) Let* $p : \mathbb{R}^d \to \mathbb{R}$

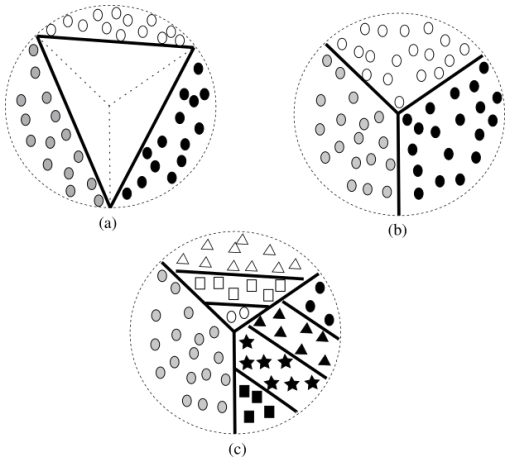

Figure 5 Three set of examples in $\mathbb{R}^2$ showing different linear separable conditions. Thick lines represent class boundaries. (a) Strongly linear separable examples with 3 classes (linearly separable in $\mathbb{R}^3$). (b) Weakly linear separable examples with 3 classes. (c) Group weakly linear separable examples with 3 groups; group 1 (white) contains 3 classes, group 2 (black) contains 4 classes, and group 3 (gray) contains 1 class.

be a multivariate polynomial. There exists $c \in \ell_2$ *such that* $p(x) = \langle c, \phi(x) \rangle_{\ell_2}$ *and* $\|c\|_{\ell_2} \leq 2^{deg(p)/2} \|p\|$.

To proof Theorem 6, we need to establish the existence of multivariate polynomials that separate one class from the other. Consider class $i \in [K]$ in group $g(i)$. Its positive example x, when compared with examples from other group $j \neq g(i)$, satisfies

$$\langle u_{g(i)}, x \rangle - \langle u_j, x \rangle = \langle u_{g(i)} - u_j, x \rangle \geq \gamma,$$

implying that all examples in class i lie in

$$R_i^+ = \bigcap_{j \neq g(i)} \{x : \langle u_{g(i)} - u_j, x \rangle \geq \gamma\},$$

while all examples in other groups lie in

$$R_i^- = \bigcup_{j \neq g(i)} \{x : \langle u_{g(i)} - u_j, x \rangle \leq -\gamma\}.$$

When comparing with other classes j in the same group $g(i)$, from Lemma 3, we know that there exists thresholds b_i and t_i that can be used to separate examples from group i, i.e., all its positive examples lie in

$$\hat{R}_i^+ = \{x : \langle u'_{g(i)}, x\rangle \geq b_i + \gamma\} \cap \{x : \langle u'_{g(i)}, x\rangle \leq t_i - \gamma\},$$

while examples from other classes in group $g(i)$ lie in

$$\hat{R}_i^- = \{x : \langle u'_{g(i)}, x\rangle \leq b_i - \gamma\} \cup \{x : \langle u'_{g(i)}, x\rangle \geq t_i + \gamma\}.$$

Let $v_b = \frac{b_i}{\|u'_{g(i)}\|} u'_{g(i)}$ and $v_t = \frac{t_i}{\|u'_{g(i)}\|} u'_{g(i)}$. Both sets can be expressed as

$$\begin{aligned}\hat{R}_i^+ = &\{x : \langle u'_{g(i)}, x\rangle \geq \langle u'_{g(i)}, v_b\rangle + \gamma\} \cap \\ &\{x : \langle u'_{g(i)}, x\rangle \leq \langle u'_{g(i)}, v_t\rangle - \gamma\},\end{aligned}$$

while examples from other classes in group $g(i)$ lie in

$$\begin{aligned}\hat{R}_i^- = &\{x : \langle u'_{g(i)}, x\rangle \leq \langle u'_{g(i)}, v_b\rangle - \gamma\} \cup \\ &\{x : \langle u'_{g(i)}, x\rangle \geq \langle u'_{g(i)}, v_t\rangle + \gamma\}.\end{aligned}$$

From Lemma 4, for class i, it is enough to establish a multivariate polynomial p_i such that

$$\begin{aligned} x \in R_i^+ \cap \hat{R}_i^+ \quad &\Rightarrow \quad p_i(x) \geq \gamma'/2, \\ x \in R_i^- \cup \hat{R}_i^- \quad &\Rightarrow \quad p_i(x) \leq -\gamma'/2. \end{aligned}$$

This is shown in the Theorem 5 below. This theorem is fairly technical and is proved in Literature 7.

Theorem 5. *(Polynomial approximation of intersection of halfspaces) Let* $v_1, v_2, \ldots, v_m \in V$ *such that* $\|v_1\|, \|v_2\|, \ldots, \|v_m\| \leq 1$. *Let* $v_b, v_t \in V$ *such that* $\|v_b\| \leq 1$ *and* $\|v_t\| \leq 1$.

Let $v' \in V$ *such that* $\|v'\| \leq 1$. *Let* $\gamma \in (0,1)$ *and* $x \in \mathrm{B}(0,1)$. *There exists a multivariate polynomial* $p : \mathbb{R}^d \to \mathbb{R}$ *such that*

1. $p(x) \geq \frac{1}{2}$ *for all* $x \in (\bigcap_{i=1}^{m} \{x : \langle v_i, x\rangle \geq \gamma\}) \cap \{x : \langle x, v'\rangle \geq \langle v_b, v'\rangle + \gamma\} \cap \{x : \langle x, v'\rangle \leq \langle v_t, v'\rangle - \gamma\}$,
2. $p(x) \leq -\frac{1}{2}$ *for all* $x \in (\bigcup_{i=1}^{m} \{x : \langle v_i, x\rangle \leq -\gamma\}) \cup \{x : \langle x, v'\rangle \leq \langle v_b, v'\rangle - \gamma\} \cup \{x : \langle x, v'\rangle \geq \langle v_t, v'\rangle + \gamma\}$,
3. $deg(p) = \lceil \log_2(2m+4) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil$,
4. $\|p\| \leq \frac{9}{2} \left[420 \lceil \log_2(2m+4) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right]^{\frac{\lceil \log_2(2m+4) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil}{2}}$

Our proof follows the approach in Beygelzimer *et al.* (2019).

Proof of Theorem 5. Let $r = \lceil \log_2(2m+4) \rceil$ and $s = \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil$. Define the polynomial $p : \mathbb{R}^d \to \mathbb{R}$ as

$$\begin{aligned} p(x) = m + \frac{5}{2} - \sum_{i=1}^{m} (T_s(1 - \langle v_i, x\rangle))^r \\ - (T_s(1 - \langle x - v_b, v'\rangle/2))^r \\ - (T_s(1 - \langle v_t - x, v'\rangle/2))^r. \end{aligned}$$

First, consider the case when

$$\begin{aligned} x \in \left(\bigcap_{i=1}^{m} \{x : \langle v_i, x\rangle \geq \gamma\} \right) \cap \{x : \langle x, v'\rangle \geq \langle v_b, v'\rangle + \gamma\} \cap \\ \{x : \langle x, v'\rangle \leq \langle v_t, v'\rangle - \gamma\}. \end{aligned}$$

Note that $\langle v_i, x\rangle \geq \gamma$ for all $i \in [m]$. Since $\|x\| \leq 1$ and $\|v_i\| \leq 1$, we have $\langle v_i, x\rangle \in [0,1]$; thus, $(T_s(1 - \langle v_i, x\rangle))^r \in [-1,1]$. Consider the terms involving v_b

and v_t. Since $\|x\|, \|v_b\|, \|v_t\| \leq 1$, we have that $\|x - v_b\| \leq 2$ and $\|v_t - x\| \leq 2$. This implies that $1 \geq \langle x - v_b, v' \rangle / 2 \geq \gamma/2$ and $1 \geq \langle v_t - x, v' \rangle / 2 \geq \gamma/2$; hence, $(T_s(1 - \langle x - v_b, v' \rangle / 2))^r \in [-1, 1]$ and $(T_s(1 - \langle v_t - x, v' \rangle / 2))^r \in [-1, 1]$. Therefore,

$$p(x) \geq m + \frac{5}{2} - m - 1 - 1 \geq \frac{1}{2}.$$

Now consider the case when

$$x \in \bigcup_{i=1}^{m} \{x : \langle v_i, x \rangle \leq -\gamma\} \cup \{x : \langle x, v' \rangle \leq \langle v_b, v' \rangle - \gamma\} \cup \{x : \langle x, v' \rangle \geq \langle v_t, v' \rangle + \gamma\}$$

There are two subcases to consider.

Subcase 1: Suppose that for some i, $\langle v_i, x \rangle \leq -\gamma$. In this case, $1 - \langle v_i, x \rangle \geq 1 + \gamma$ and Lemma 1 (part 6) implies that

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

and thus, $(T_s(1 - \langle v_i, x \rangle))^r \geq 2^r \geq 2m + 4$.

Since $T_s(1 - \langle v_i, x \rangle))^r \geq -1$ for all i, $(T_s(1 - \langle x - v_b, v' \rangle / 2))^r \geq -1$, and $(T_s(1 - \langle v_t - x, v' \rangle / 2))^r \geq -1$, we have that

$$\begin{aligned} p(x) &= m + \frac{5}{2} - (T_s(1 - \langle v_i, x \rangle))^r \\ &\quad - \sum_{j \in [m] j \neq i} (T_s(1 - \langle v_j, x \rangle))^r \\ &\quad - (T_s(1 - \langle x - v_b, v' \rangle / 2))^r \\ &\quad - (T_s(1 - \langle v_t - x, v' \rangle / 2))^r \\ &\leq m + \frac{5}{2} - (2m + 4) + (m - 1) + 2 \leq -\frac{1}{2}. \end{aligned}$$

Subcase 2: Consider the other case when for all i, $\langle v_i, x\rangle > -\gamma$. We deal with the case that $\langle x, v'\rangle \leq \langle v_b, v'\rangle - \gamma$. The case when $\langle x, v'\rangle \geq \langle v_t, v'\rangle + \gamma$ can be handled similarly.

Since $\langle x - v_b, v'\rangle \leq -\gamma$, we have $1 - \langle x - v_b, v'\rangle/2 \geq 1 + \gamma/2$. Lemma 1 (part 6) implies that

$$T_s(1 - \langle x - v_b, v'\rangle/2) \geq 1 + s^2\gamma/2 \geq 1 + 2/2 \geq 2,$$

and $(T_s(1 - \langle x - v_b, v'\rangle/2))^r \geq 2m + 4$. Applying the same argument as in Subcase 1, this implies that $p(x) \leq -\frac{1}{2}$.

The degree of p is the maximum degree of the terms $(T_s(1 - \langle v_i, x\rangle))^r$, $(T_s(1 - \langle x - v_b, v'\rangle/2))^r$, and $(T_s(1 - \langle v_t - x, v'\rangle)/2)^r$; thus, it is $r \cdot s$.

Finally, we prove the upper bound of norm of p. Let $f_i(x) = 1 - \langle v_i, x\rangle$, let $k_b(x) = 1 - \langle x - v_b, v'\rangle/2$, $k_t(x) = 1 - \langle v_t - x, v'\rangle/2$.

$$\|f_i\|^2 = 1 + \|v_i\|^2 \leq 1 + 1 = 2,$$

$$\|k_b\|^2 = 1 + \frac{\|x - v_b\|^2 \cdot \|u'\|^2}{2} \leq 1 + \frac{4 \cdot 1}{2} = 3$$

and

$$\|k_t\|^2 = 1 + \frac{\|v_t - x\|^2 \cdot \|u'\|^2}{2} \leq 1 + \frac{4 \cdot 1}{2} = 3.$$

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

We first deal with $\|T_s(1 - \langle v_i, x\rangle)\|^2$. By lemma 1 and 2, $s + 1 \leq 2^s$ for any

non-negative integer, we have

$$\begin{aligned}
\|T_s(1-\langle v_i,x\rangle)\|^2 &= \|T_s(f_i)\|^2 \\
&= \left\|\sum_{j=0}^{s} c_j (f_i)^j\right\|^2 \\
&\le (s+1)\sum_{j=0}^{s} \left\|c_j (f_i)^j\right\|^2 \\
&= (s+1)\sum_{j=0}^{s} c_j^2 \left\|(f_i)^j\right\|^2 \\
&\le (s+1)\sum_{j=0}^{s} c_j^2 j^j \left\|f_i\right\|^{2j} \\
&\le (s+1)\sum_{j=0}^{s} c_j^2 j^j 2^{2j} \\
&\le (s+1)s^s 2^{2s}\sum_{j=0}^{s} c_j^2 \\
&= (s+1)s^s 2^{2s}\|T_s\|^2 \\
&= (s+1)s^s 2^{2s}(1+\sqrt{2})^{2s} \\
&= (s+1)\left(4(1+\sqrt{2})^2 s\right)^s \\
&\le (8(1+\sqrt{2})^2 s)^s \\
&\le (47s)^s.
\end{aligned}$$

The other two terms $\|T_s(\langle x-v_b,v'\rangle/2)\|^2$ and $\|T_s(\langle v_t-x,v'\rangle/2)\|^2$ can be

analyzed similarly. We have that

$$\begin{aligned}
\|T_s(\langle x - v_b, v' \rangle/2)\|^2 &= \|T_s(k_b)\|^2 \\
&= \left\| \sum_{j=0}^{s} c_j (k_b)^j \right\|^2 \\
&\leq (s+1) \sum_{j=0}^{s} c_j^2 j^j \|k_b\|^{2j} \\
&\leq (s+1) \sum_{j=0}^{s} c_j^2 j^j 3^{2j} \\
&\leq (s+1) s^s 3^{2s} \sum_{j=0}^{s} c_j^2 \\
&= (s+1) s^s 3^{2s} \|T_s\|^2 \\
&= (s+1) s^s 3^{2s} (1+\sqrt{2})^{2s} \\
&= (s+1) \left(9(1+\sqrt{2})^2 s \right)^s \\
&\leq (9(1+\sqrt{2})^2 s)^s \\
&\leq (105s)^s
\end{aligned}$$

and

$$\begin{aligned}
\|T_s(\langle v_t - x, v' \rangle/2)\|^2 &= \|T_s(k_t)\|^2 \\
&= \left\| \sum_{j=0}^{s} c_j (k_t)^j \right\|^2 \\
&\leq (105s)^s.
\end{aligned}$$

Finally,

$$\begin{aligned}
\|p\| &\leq m + \frac{5}{2} + \sum_{i=1}^{m} \|T_s(f_i)^r\| + \|T_s(k_b)^r\| + \|T_s(k_t)^r\| \\
&= m + \frac{5}{2} + \sum_{i=1}^{m} \sqrt{\|T_s(f_i)^r\|^2} \\
&\quad + \sqrt{\|T_s(k_b)^r\|^2} + \sqrt{\|T_s(k_t)^r\|^2} \\
&\leq m + \frac{5}{2} + \sum_{i=1}^{m} \sqrt{r^{rs} \|T_s(f_i)^r\|^{2r}} \\
&\quad + \sqrt{r^{rs} \|T_s(k_b)^r\|^{2r}} + \sqrt{r^{rs} \|T_s(k_t)^r\|^{2r}} \\
&\leq m + \frac{5}{2} + mr^{rs/2}(47s)^{rs/2} + r^{rs/2}(105s)^{rs/2} \\
&\quad + r^{rs/2}(105s)^{rs/2} \\
&\leq m + \frac{5}{2} + (m+2)(105rs)^{rs/2}.
\end{aligned}$$

Using the fact that $m \leq \frac{1}{2}2^r$ and $r, s \geq 1$, we then have

$$\begin{aligned}
\|p\| &\leq m + \frac{5}{2} + (m+2)(105rs)^{rs/2} \\
&\leq \frac{1}{2}2^r + \frac{5}{2} + \left(\frac{1}{2}2^r + 2\right)(105rs)^{rs/2} \\
&\leq 2 \cdot 2^r + \frac{5}{2} \cdot 2^r (105rs)^{rs/2} \\
&= 2^r \left(2 + \frac{5}{2}\right)(105rs)^{rs/2} \\
&\leq 4^{rs/2} \cdot \frac{9}{2}(105rs)^{rs/2} \\
&= \frac{9}{2}(420rs)^{rs/2}.
\end{aligned}$$

Substitutions of r and s finish the proof. □

Our main technical result is the following margin transformation using the rational kernel.

Theorem 6. *(Margin transformation). Let* $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T) \in \mathrm{B}(0,1) \times [K]$ *be a sequence of labeled examples that is group weakly linear separable with mar-*

gin $\gamma > 0$. Let L be number of group weakly separable such that $L \leq K$. Let ϕ defined as in (1) let

$$\gamma' = \frac{\left[840\lceil \log_2(2L+2)\rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right]^{-\frac{\lceil \log_2(2L+2)\rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil}{2}}}{9\sqrt{L}},$$

The feature map ϕ makes the sequence $(\phi(x_1), y_1), (\phi(x_2), y_2), \ldots, (\phi(x_T), y_T)$ strongly linearly separable with margin γ'.

We note that the margin depends on L, the number of groups, instead of K, the number of classes. Using Theorem 6 with Theorem 3 we obtain the following mistake bound for our algorithm.

Proof of Theorem 6. Consider class $i \in [K]$. We will apply Theorem 5. For $j \in \{1, \ldots, L-1\}$, let

$$v_j = \begin{cases} u_{g(i)} - u_j, & \text{if } j < g(i), \\ u_{g(i)} - u_{j+1}, & \text{if } j > g(i). \end{cases}$$

Also, let $v' = u'_{g(i)}$, $v_b = \frac{b_i}{\|u'_{g(i)}\|} u'_{g(i)}$ and $v_t = \frac{t_i}{\|u'_{g(i)}\|} u'_{g(i)}$.

From Theorem 5, there exists a multivariate polynomial $p_i : \mathbb{R}^d \to \mathbb{R}$ such that for all $t \in [T]$ and the sequence $(x_1, y_1), (x_2, y_2), (x_t, y_t), \ldots, (x_T, y_T)$, we have

- if $y_t = i$, $p_i(x_t) \geq \frac{1}{2}$, since $x_t \in R_i^+ \cap \hat{R}_i^+$, and
- if $y_t \neq i$, $p_i(x_t) \leq -\frac{1}{2}$, since $x_t \in R_i^- \cap \hat{R}_i^-$.

It is left to check the properties of p. Theorem 5 implies that

$$\|p\| \leq \frac{9}{2} \left[420\lceil \log_2(2L+2)\rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right]^{\frac{\lceil \log_2(2L+2)\rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil}{2}}$$

By Lemma 4, there exists $c_i \in \ell_2$ such that $\langle c_i, \phi(x) \rangle = p_i(x)$, and

$$\|c_i\|_{\ell_2} \leq \frac{9}{2} \left[840 \lceil \log_2(2L+2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right]^{\frac{\lceil \log_2(2L+2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil}{2}}.$$

We are ready to construct strongly separable vectors for our group weakly separable case in ℓ_2 such that $\|z_1\|^2 + \|z_2\|^2 + \ldots + \|z_L\|^2 \leq 1$ and for all $t \in [T]$, $\langle z_{y_t}, x_t \rangle \geq \gamma$, and for all $j \neq y_t$, $\langle z_j, x_t \rangle \leq -\gamma$, by scaling c_i appropriately as follows. We can let

$$z_i = \frac{c_i}{\sqrt{L} \cdot \frac{9}{2} \left[840 \lceil \log_2(2L+2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right]^{\frac{\lceil \log_2(2L+2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil}{2}}},$$

and

$$\gamma = \frac{\left[840 \lceil \log_2(2L+2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil \right]^{-\frac{\lceil \log_2(2L+2) \rceil \cdot \left\lceil \sqrt{\frac{2}{\gamma}} \right\rceil}{2}}}{9\sqrt{L}},$$

then the theorem follows. □

Corollary 2. *(Mistake bound for group weakly linearly separable case) Let K be positive integer, $L \leq K$ and γ be positive real number. The mistake bound made by Algorithm 4 when the examples are group weakly linearly separable with margin γ with L groups is at most $K \cdot 2^{\tilde{O}(\sqrt{1/\gamma} \log L)}$.*

Note that multiplicative factor of K is hidden from the second bound of Beygelzimer *et al.* (2019) because of the $\tilde{O}$ notation on the exponent. We cannot do that because in our exponent we have only $\log L$ which can be much smaller than K. Their actual bound (showing K), which can be compared to ours, is $K \cdot 2^{\tilde{O}(\sqrt{1/\gamma} \log K)}$.

EXPERIMENTS

While we focus mostly on the theoretical aspect of the problem, we performed some experiment to visualize the algorithm. We generated a dataset in $\mathbb{R}^2$ under the group weakly linear separable condition, with $K = 9$ classes and $L = 3$ groups with margin γ, shown in Fig 6.

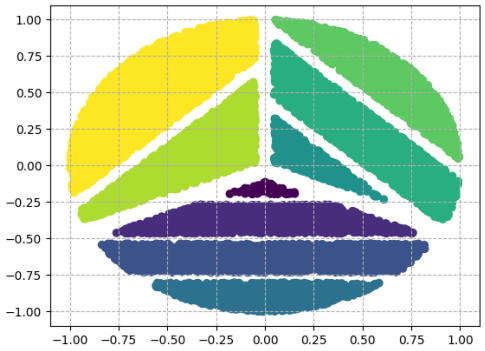

Figure 6 Group weakly separable dataset in $\mathbb{R}^2$.

We compared two versions of the bandit multiclass perceptron Beygelzimer *et al.* (2019), the standard one and the kernelized one (using the rational kernel). Since the standard one only works with strongly separable case, it would definitely fail in this experiment, but we used it to give an overall sense of improvement for the kernelized version. We ran both algorithms for 10^5 steps for 5 times. Fig. 7 shows the result. The kernelized version made on average 15831 mistakes (15.8%), while the standard one made on average 76759.4 mistakes (76.7%). Theoretically, the kernelized version should stop making mistakes at some point, but since the number of steps that we ran is too low, we can only see that increasing rate of the number of mistakes decreases over time.

To see the decision boundary, we ploted the contours of the corresponding polynomials for two classes shown in Fig. 8 and Fig. 9. Note that the class in Fig. 9 was much harder to learn as its boundary still overlapped with other classes (i.e., mistakes

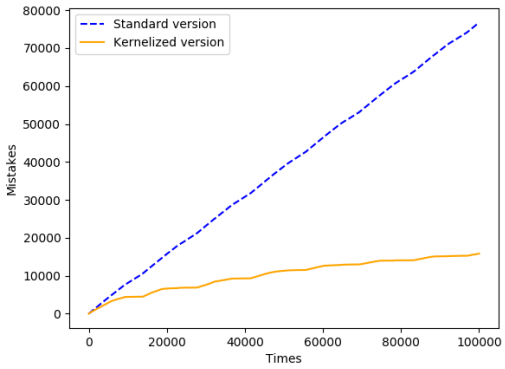

Figure 7 Comparison of the standard algorithm and the kernelized algorithm with $T = 10^5$.

could still be made).

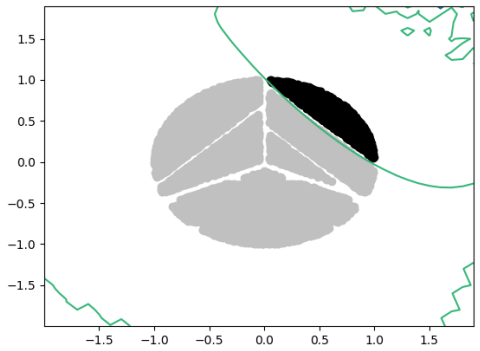

Figure 8 The boundary line of a class (in black) of the kernelized algorithm after $T = 10^5$ steps.

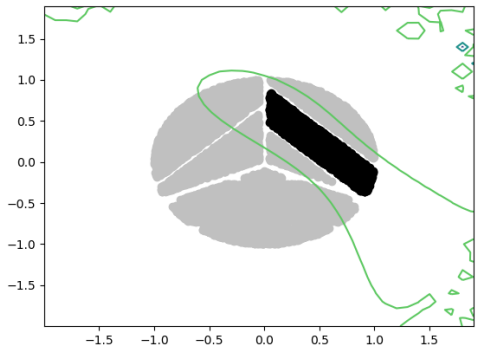

Figure 9 The boundary line of a class (in black) of the kernelized algorithm after $T = 10^5$ steps.

LITERATURE CITED

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

Beygelzimer, A., D. Pal, B. Szorenyi, D. Thiruvenkatachari, C.-Y. Wei and C. Zhang. 2019. Bandit multiclass linear classification: Efficient algorithms for the separable case, pp. 624–633. In Chaudhuri, K. and R. Salakhutdinov, eds. **Proceedings of the 36th International Conference on Machine Learning**. PMLR, Long Beach, California, USA.

Chen, G., G. Chen, J. Zhang, S. Chen and C. Zhang. 2009. Beyond banditron: A conservative and efficient reduction for online multiclass prediction with bandit setting model, pp. 71–80. In **ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009**.

Crammer, K. and Y. Singer. 2003. Ultraconservative online algorithms for multiclass problems. **J. Mach. Learn. Res.** 3: 951–991.

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

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

Mason, J. and D. Handscomb. 2002. **Chebyshev Polynomials**. CRC Press.

Rosenblatt, F. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**. 65 (6): 386–408.

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

Shawe-Taylor, J. and N. Cristianini. 2004. **Kernel Methods for Pattern Analysis**. Cambridge University Press, New York, NY, USA.