ON ACCELERATED PERCEPTRONS AND BEYOND

Guanghui Wang[1], **Rafael Hanashiro**[2], **Etash Guha**[1], **Jacob Abernethy**[1,3]
[1]College of Computing, Georgia Tech, Atlanta, GA, USA
[2]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA
[3]Google Research, Atlanta, GA 30309
{gwang369,etash}@gatech.edu, rafah@mit.edu, abernethyj@google.com

ABSTRACT

The classical Perceptron algorithm of Rosenblatt can be used to find a linear threshold function to correctly classify n linearly separable data points, assuming the classes are separated by some margin $\gamma > 0$. A foundational result is that Perceptron converges after $O(1/\gamma^2)$ iterations. There have been several recent works that managed to improve this rate by a quadratic factor, to $O(\sqrt{\log n}/\gamma)$, with more sophisticated algorithms. In this paper, we unify these existing results under one framework by showing that they can all be described through the lens of solving min-max problems using modern acceleration techniques, mainly through *optimistic* online learning. We then show that the proposed framework also leads to improved results for a series of problems beyond the standard Perceptron setting. Specifically, a) for the *margin maximization* problem, we improve the state-of-the-art result from $O(\log t/t^2)$ to $O(1/t^2)$, where t is the number of iterations; b) we provide the first result on identifying the *implicit bias* property of the classical Nesterov's accelerated gradient descent (NAG) algorithm, and show NAG can maximize the margin with an $O(1/t^2)$ rate; c) for the classical p*-norm Perceptron* problem, we provide an algorithm with $O(\sqrt{(p-1)\log n}/\gamma)$ convergence rate, while existing algorithms suffer the $O((p-1)/\gamma^2)$ convergence rate.

1 INTRODUCTION

In this paper, we revisit the problem of learning a linear classifier, which is one of the most important and fundamental tasks of machine learning (Bishop, 2007). In this problem, we are given a set $\mathcal{S}$ of n training examples, and the goal is to find a linear classifier that correctly separates $\mathcal{S}$ as fast as possible. The most well-known algorithm is Perceptron (Rosenblatt, 1958), which can converge to a perfect (mistake-free) classifier after $\Omega(1/\gamma^2)$ number of iterations, provided the data is linearly separable with some margin $\gamma > 0$ (Novikoff, 1962). Over subsequent decades, many variants of Perceptron have been developed (Aizerman, 1964; Littlestone, 1988; Wendemuth, 1995; Freund & Schapire, 1999; Cesa-Bianchi et al., 2005, to name a few). However, somewhat surprisingly, there has been little progress in substantially improving the fundamental Perceptron iteration bound presented by Novikoff (1962). It is only recently that a number of researchers have discovered *accelerated* variants of the Perceptron with a faster $\Omega(\sqrt{\log n}/\gamma)$ iteration complexity, although with a slower per-iteration cost. These works model the problem in different ways, e.g., as a non-smooth optimization problem or an empirical risk minimization task, and they have established faster rates using sophisticated optimization tools. Soheili & Pena (2012) put forward the *smooth Perceptron*, framing the objective as a non-smooth strongly-concave maximization and then applying *Nestrov's excessive gap technique* (NEG, Nesterov, 2005). Yu et al. (2014) proposed the *accelerated Perceptron* by furnishing a convex-concave objective that can be solved via the mirror-prox method (Nemirovski, 2004). Ji et al. (2021) put forward a third interpretation, obtaining the accelerated rate by minimizing the empirical risk under exponential loss with a momentum-based normalized gradient descent algorithm.

Following this line of research, in this paper, we present a *unified* analysis framework that reveals the exact relationship among these methods that share the same order of convergence rate. Moreover, we show that the proposed framework also leads to improved results for various problems beyond the standard Perceptron setting. Specifically, we consider a general zero-sum game that involves

two players (Abernethy et al., 2018): a main player that chooses the classifier, and an auxiliary player that picks a distribution over data. The two players compete with each other by performing *no-regret online learning* algorithms (Hazan, 2016; Orabona, 2019), and the goal is to find the equilibrium of some convex-concave function. We show that, under this dynamic, all of the existing accelerated Perceptrons can find their equivalent forms. In particular, these perceptrons can be described as a dynamic where two players solving the game via performing *optimistic online learning* strategies (Rakhlin & Sridharan, 2013), which is one of the most important classes of algorithms in online learning. Note that implementing online learning algorithms (even optimistic strategies) to solve zero-sum games has already been extensively explored (e.g., Rakhlin & Sridharan, 2013; Daskalakis et al., 2018; Wang & Abernethy, 2018; Daskalakis & Panageas, 2019). However, we emphasize that our main novelty lies in showing that all of the existing accelerated Perceptrons, developed with advanced algorithms from different areas, can be perfectly described under this unified framework. It greatly simplifies the analysis of accelerated Perceptrons, as their convergence rates can now be easily obtained by plugging-in off-the-shelf regret bounds of optimistic online learning algorithms. Moreover, the unified framework reveals a close connection between the smooth Perceptron and the accelerated Perceptron of Ji et al. (2021):

Theorem 1 (informal)**.** *Smooth Perceptron and the accelerated Perceptron of Ji et al. (2021) can be described as a dynamic where the two players employ the optimistic-follow-the-regularized-leader (OFTRL) algorithm to play. The main difference is that the smooth Perceptron outputs the* weighted average *of the main player's historical decisions, while the accelerated Perceptron of Ji et al. (2021) outputs the* weighted sum*.*

Beyond providing a deeper understanding of accelerated Perceptrons, our framework also provides improved new results for several other important areas:

- *Implicit bias* analysis. The seminal work of Soudry et al. (2018) shows that, for linearly separable data, minimizing the empirical risk with the vanilla gradient descent (GD) gives a classifier which not only has zero training error (thus can be used for linear separation), but also *maximizes the margin*. This phenomenon characterizes the *implicit bias* of GD, as it implicitly prefers the (ℓ_2-)maximal margin classifier among all classifiers with a positive margin, and analysing the implicit bias has become an important tool for understanding why *classical* optimization methods generalize well for supervised machine learning problems. Soudry et al. (2018) show that GD can maximize the margin in an $O(1/\log t)$ rate, and this is later improved to $O(1/\sqrt{t})$ (Nacson et al., 2019) and then $O(1/t)$ (Ji & Telgarsky, 2021). The state-of-the-art algorithm is proposed by Ji et al. (2021), who show that their proposed momentum-based GD has an $O(\log t/t^2)$ margin-maximization rate. In this paper, we make two contributions toward this direction:
  1. We show that, under our analysis framework, we can easily improve the margin maximization rate of the algorithm of Ji et al. (2021) from $O(\log t/t^2)$ to $O(1/t^2)$;
  2. Although previous work has analyzed the implicit bias of GD and momentum-based GD, it is still unclear how the classical Nesterov's accelerated gradient descent (NAG, Nesterov, 1988) will affect the implicit bias. In this paper, through our framework, we show that NAG with appropriately chosen parameters also enjoys an $O(1/t^2)$ margin-maximization rate. To our knowledge, it is the first time the implicit bias property of NAG is proved.
- p-norm Perceptron. Traditional work on Perceptrons typically assumes the feature vectors lie in an ℓ_2-ball. A more generalized setting is considered in Gentile (2000), who assumes the feature vectors lie inside an ℓ_p-ball, with $p \in [2, \infty)$. Their proposed algorithm requires $O(p/\gamma^2)$ number of iterations to find a zero-error classifier. In this paper, we develop a new Perceptron algorithm for this problem under our framework based on optimistic strategies, showing that it enjoys an accelerated $O(\sqrt{p \log n}/\gamma)$ rate.

2 Related Work

This section briefly reviews the related work on Perceptron algorithms, implicit-bias analysis, and game theory. The background knowledge on (optimistic) online learning is presented in the Preliminaries (Section 3).

Accelerated Perceptrons and p-norm Perceptron The study of Perceptron algorithms has an extensive history dating back to the mid-twentieth century (Rosenblatt, 1958; Novikoff, 1962). However, it is only recently that progress on improving the fundamental $\Omega(1/\gamma^2)$ iteration bound of the vanilla Perceptron in the standard setting has been made. Specifically, the smooth Perceptron proposed by Soheili & Pena (2012) achieves an $\Omega(\sqrt{\log n}/\gamma)$ rate by maximizing an auxiliary non-smooth strongly-concave function with NEG (Nesterov, 2005). The same accelerated rate is later obtained by two other work (Yu et al., 2014; Ji et al., 2021). The former considers a bi-linear saddle point problem and employs the mirror-prox method (Nemirovski, 2004). The latter applies momentum-based GD to minimize the empirical risk with exponential loss. In this paper, we show that these algorithms can be unitedly analysed under our framework.

Apart from the above accelerated Perceptrons, there exists another class of algorithms which enjoy an $O(\mathrm{poly}(n)\log(1/\gamma))$ convergence rate (Dunagan & Vempala, 2004; Peña & Soheili, 2016; Dadush et al., 2020). These methods typically call (accelerated) Perceptrons as a subroutine, and then apply the re-scaling technique to adjust the decision periodically. Although the dependence on γ becomes better, the polynomial dependence on n of these methods makes them computationally inefficient for large-scale data sets. In this paper, we focus on accelerated Perceptrons with $O(\sqrt{\log n}/\gamma)$ rate, and leave explaining the re-scaling type algorithms as future work.

The p-norm Perceptron problem (Gentile & Littlestone, 1999) is a natural extension of the classical Perceptron setting, which assumes the p-norm of the feature vectors are bounded, where $2 \leq p < \infty$. Gentile (2001) shows that a mirror-descent-style update guarantees an $O(p/\gamma^2)$ convergence rate. By contrast, our proposed algorithm achieves a tighter $O(\sqrt{p\log n}/\gamma)$ convergence rate.

Implicit-bias analysis In many real-world applications, directly minimizing the empirical risk (without any regularization) by first-order methods can provide a model which, not only enjoys low training error, but also generalizes well (Soudry et al., 2018). This is usually considered as the *implicit bias introduced by the optimization methods* (Soudry et al., 2018). Explaining this phenomenon is a crucial step towards understanding the generalization ability of *commonly-used* optimization methods. For linear separable data, Soudry et al. (2018) proves that, when minimizing the empirical risk with exponential loss, the vanilla GD can maximize the margin in an $O(1/\log t)$ rate. This result implies the implicit bias of GD towards the ℓ_2-maximal margin classifier. Later, Nacson et al. (2019) show that GD with a function-value-dependant decreasing step-size enjoys an $O(1/\sqrt{t})$ margin-maximization rate. Ji & Telgarsky (2021) improve this result to $O(1/t)$ by employing a faster-decreasing step size. Ji et al. (2021) design a momentum-based GD that maximizes the margin with an $O(\log t/t^2)$ rate. However, it remained unclear whether this rate could be further improved and how to analyze the margin maximization ability of other classical optimization methods such as NAG. In this paper, we provide positive answers to both questions. Fianlly, we note that (Ramdas & Pena, 2016) prove a varint of the Perceptron algorithm named normalized perceptron can also converge to a maximum margin classifer in an $O(1/\sqrt{t})$ convergence rate.

Games and no-regret dynamics Our framework is motivated by the line of research that links optimization methods for convex optimization to equilibrium computation with no-regret dynamics. The seminal work of Rakhlin & Sridharan (2013) recovers the classical mirror-prox method (Nemirovski, 2004) with Optimistic online mirror descent. Abernethy & Wang (2017) show that the well-known Frank-Wolfe algorithm can be seen as applying (optimistic) online algorithms to compute the equilibrium of a special zero-sum game called *Fenchel game*, which is constructed via the Fenchel duality of the objective function. Later, researchers demonstrate that other classical optimization methods for smooth optimization, such as Heavy-ball and NAG, can also be described similarly (Abernethy et al., 2018; Wang & Abernethy, 2018; Wang et al., 2021). We highlight the differences between this work and the previous ones: 1) Our analysis does not involve the Fenchel game or Fenchel duality; instead, we directly work on the (regularized) min-max game designed for linear classification. 2) Most of the previous work mainly focus on understanding optimization algorithms for *smooth* optimization, and it was unclear how to understand algorithm such as NEG under the game framework. 3) Although both Abernethy et al. (2018) and our work analyze NAG, the goals are significantly different: Abernethy et al. (2018) focus on the optimization problem itself, while we consider how minimizing the empirical risk would affect the *margin*. 4) To our knowledge, the link between the implicit-bias problems and no-regret dynamics is also new.

3 PRELIMINARIES

Notation. We use lower case bold face letters $\mathbf{x}, \mathbf{y}$ to denote vectors, lower case letters a, b to denote scalars, and upper case letters A, B to denote matrices. For a vector $\mathbf{x} \in \mathbb{R}^d$, we use x_i to denote the i-th component of $\mathbf{x}$. For a matrix $A \in \mathbb{R}^{n \times d}$, let $A_{(i,:)}$ be its i-th row, $A_{(:,j)}$ the j-th column, and $A_{(i,j)}$ the i-th element of the j-th column. We use $\|\cdot\|$ to denote a general norm, $\|\cdot\|_*$ its dual norm, and $\|\cdot\|_p$ the ℓ_p-norm. For a positive integer n, we denote the set $\{1, \ldots, n\}$ as $[n]$, and the n-dimensional simplex as Δ^n. Let $E : \Delta^n \mapsto \mathbb{R}$ be the negative entropy function, defined as $E(\mathbf{p}) = \sum_{i=1}^n p_i \log p_i, \forall \mathbf{p} \in \Delta^n$. For some strongly convex function $R(\mathbf{w}) : \mathcal{W} \mapsto \mathbb{R}$, define the Bregman divergence between any two points $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$ as:

$$D_R(\mathbf{w}, \mathbf{w}') = R(\mathbf{w}) - R(\mathbf{w}') - (\mathbf{w} - \mathbf{w}')^\top \nabla R(\mathbf{w}').$$

Online convex optimization (OCO) Here we review a general *weighted* OCO framework, proposed by Abernethy et al. (2018). In each round $t = 1, \ldots, T$ of this paradigm, a learner first chooses a decision $\mathbf{z}_t$ from a convex set $\mathcal{Z} \subseteq \mathbb{R}^d$, then observes a loss function $f_t(\cdot) : \mathcal{Z} \to \mathbb{R}$ as well as a weight $\alpha_t > 0$, and finally updates the decision. The performance of the learner is measured by the *weighted* regret, which is defined as $R_T = \sum_{t=1}^T \alpha_t f_t(\mathbf{z}_t) - \min_{\mathbf{z} \in \mathcal{Z}} \sum_{t=1}^T \alpha_t f_t(\mathbf{z})$.

Perhaps the most natural method for OCO is follow-the-leader (FTL), which simply picks the empirically best decision at each round: $\mathbf{z}_t = \operatorname{argmin}_{\mathbf{z} \in \mathcal{Z}} \sum_{i=1}^{t-1} \alpha_i f_i(\mathbf{z})$. However, FTL is unstable, and one can easily find counter-examples where FTL suffers linear regret (Shalev-Shwartz, 2011). A classical way to address this limitation is by adding a regularizer to the objective: $\mathbf{z}_t = \operatorname{argmin}_{\mathbf{z} \in \mathcal{Z}} \eta \sum_{i=1}^{t-1} \alpha_i f_i(\mathbf{z}) + D_R(\mathbf{z}, \mathbf{z}_0)$, where $\eta > 0$ is the step size, and $\mathbf{z}_0$ is the initial decision. This method is called follow-the-regularized-leader (Hazan, 2016), and it can achieve a sub-linear regret bound with appropriately chosen η. Moreover, tighter bounds are also possible in favored cases with more advanced techniques. In this paper, we consider FTRL equipped with optimistic strategies (i.e., Optimistic FTRL, Rakhlin & Sridharan, 2013; Orabona, 2019), given by

$$\text{OFTRL}[R, \mathbf{z}_0, \psi_t, \eta, \mathcal{Z}] : \mathbf{z}_t = \operatorname{argmin}_{\mathbf{z} \in \mathcal{Z}} \eta \left[\sum_{i=1}^{t-1} \alpha_i f_i(\mathbf{z}) + \alpha_t \psi_t(\mathbf{z}) \right] + D_R(\mathbf{z}, \mathbf{z}_0),$$

where an additional function ψ_t is added, which is an approximation of the next loss f_t. A tighter regret bound can be achieved when ψ_t is close enough to f_t. Next, we introduce two special cases of OFTRL:

$$\text{OFTL}[\psi_t, \mathcal{Z}] : \mathbf{z}_t = \operatorname*{argmin}_{\mathbf{z} \in \mathcal{Z}} \sum_{j=1}^{t-1} \alpha_j f_j(\mathbf{z}) + \alpha_t \psi_t(\mathbf{z})$$

$$\text{FTRL}^+[R, \mathbf{z}_0, \eta, \mathcal{Z}] : \mathbf{z}_t = \operatorname*{argmin}_{\mathbf{z} \in \mathcal{Z}} \eta \cdot \sum_{j=1}^{t} \alpha_j f_j(\mathbf{z}) + \mathcal{D}_R(\mathbf{z}, \mathbf{z}_0).$$

The first one is Optimistic FTL, where the regularizer is set to be zero. The second algorithm is FTRL$^+$, which uses f_t as the optimistic function ψ_t. Finally, we note that, apart from FTL-type algorithms, there also exists optimistic methods that are developed based on mirror descent, such as optimistic online mirror decent (OMD, Rakhlin & Sridharan, 2013). Due to page limitation, the details of OMD and the regret bounds of these OCO algorithms are postponed to the Appendix A.

No-regret dynamics for zero-sum game Finally, we introduce the framework for using no-regret online algorithms to solve a zero-sum game. Consider the following general two-player game:

$$\max_{\mathbf{w} \in \mathcal{W}} \min_{\mathbf{p} \in \mathcal{Q}} g(\mathbf{w}, \mathbf{p}), \tag{1}$$

where $g(\mathbf{w}, \mathbf{p})$ is a concave-convex function, and $\mathcal{W}$ and $\mathcal{Q}$ are convex sets. The no-regret framework for solving (1) is presented in Protocol 1. In each round t of this procedure, the $\mathbf{w}$-player first picks a decision $\mathbf{w}_t \in \mathcal{W}$. Then, the $\mathbf{p}$-player observes its loss $\ell_t(\cdot) = g(\mathbf{w}_t, \cdot)$ as well as a weight α_t. After that, the $\mathbf{p}$-player picks the decision $\mathbf{p}_t$, and passes it to the $\mathbf{w}$-player. As a consequence, the $\mathbf{w}$-player observes its loss $h_t(\cdot) = -g(\cdot, \mathbf{p}_t)$ and weight α_t. Note that both ℓ_t and h_t are convex. Denote the weighted regret bounds of the two players as $R^{\mathbf{w}}$ and $R^{\mathbf{p}}$, respectively. Then, we have the following classical conclusion. The proof is shown in Appendix B.

Theorem 2. *Define* $m(\mathbf{w}) = \min_{\mathbf{p} \in \mathcal{Q}} g(\mathbf{w}, \mathbf{p})$, *and* $\overline{\mathbf{w}}$ *the weighted average of* $\{\mathbf{w}_t\}_{t=1}^T$. *Then we have that* $\forall \mathbf{w} \in \mathcal{W}$, $m(\mathbf{w}) - m(\overline{\mathbf{w}}_T) \leq \left(\sum_{t=1}^T \alpha_t\right)^{-1} \left(R^{\mathbf{p}} + R^{\mathbf{w}}\right)$.

Protocol 1 No-regret dynamics with weighted OCO

Input: $\{\alpha_t\}_{t=1}^T, \mathbf{w}_0, \mathbf{p}_0$.
Input: $\mathsf{OL}^{\mathbf{w}}, \mathsf{OL}^{\mathbf{p}}$. // The online algorithms for choosing $\mathbf{w}$ and $\mathbf{p}$.
for $t = 1, \dots, T$ **do**
  $\mathbf{w}_t \leftarrow \mathsf{OL}^{\mathbf{w}}$;
  $\mathsf{OL}^{\mathbf{p}} \leftarrow \alpha_t, \ell_t(\cdot)$; // Define $\ell_t(\cdot) = g(\mathbf{w}_t, \cdot)$
  $\mathbf{p}_t \leftarrow \mathsf{OL}^{\mathbf{p}}$;
  $\mathsf{OL}^{\mathbf{w}} \leftarrow \alpha_t, h_t(\cdot)$; // Define $h_t(\cdot) = -g(\cdot, \mathbf{p}_t)$
end for
Output: $\overline{\mathbf{w}}_T = \frac{1}{\sum_{t=1}^T \alpha_t} \sum_{t=1}^T \alpha_t \mathbf{w}_t$.

4 Understanding Existing Accelerated Perceptrons

In this section, we first introduce our main topic, i.e., the binary linear classification problem, and then present the three accelerated Perceptrons and their equivalent forms under Protocol 1. For clarity, we use $\mathbf{v}_t$ to denote the classifier updates in the original algorithms, and $\mathbf{w}_t$ the corresponding updates in the equivalent forms under Protocol 1.

4.1 Binary Linear classification and Perceptron

We focus on the binary linear classification problem, which dates back to the pioneering work of Rosenblatt (1958). Let $\mathcal{S} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^n$ be a linear-separable set of n training examples, where $\mathbf{x}^{(i)} \in \mathbb{R}^d$ is the feature vector of the i-th example, and $y^{(i)} \in \{-1, 1\}$ is the corresponding label. The goal is to efficiently find a linear classifier $\mathbf{w} \in \mathbb{R}^d$ that correctly separates all data points. More formally, let $A = [y^{(1)}\mathbf{x}^{(1)}, \dots, y^{(n)}\mathbf{x}^{(n)}]^\top$ be the matrix that contains all of the data. Then we would like to find a $\mathbf{w} \in \mathbb{R}^d$ such that

$$\min_{i \in [n]} A_{(i,:)}\mathbf{w} > 0. \tag{2}$$

This goal can be reformulated as a min-max optimization problem:

$$\max_{\mathbf{w} \in \mathbb{R}^d} \min_{i \in [n]} A_{(i,:)}\mathbf{w} = \max_{\mathbf{w} \in \mathbb{R}^d} \min_{\mathbf{p} \in \Delta^n} \mathbf{p}^\top A \mathbf{w}, \tag{3}$$

where, at the RHS of (3), $\min_{i \in [n]} A_{(i,:)}\mathbf{w}$ is rewritten as $\min_{\mathbf{p} \in \Delta^n} \mathbf{p}^\top A\mathbf{w}$. The two expressions are equivalent since the optimal distribution $\mathbf{p} \in \Delta^n$ will always put all weight on *one* training example (i.e., one row) in A. For any $\mathbf{w} \in \mathbb{R}^d$, let $\gamma(\mathbf{w}) = \min_{\mathbf{p} \in \Delta^n} \mathbf{p}^\top A\mathbf{w}$ be the *margin* of $\mathbf{w}$, and we introduce the following standard assumption.

Assumption 1. *We assume that feature vectors are bounded, i.e.,* $\|\mathbf{x}^{(i)}\|_2 \le 1$ *for all* $i \in [n]$*, and that there exists a* $\mathbf{w}^* \in \mathbb{R}^d$*, such that* $\|\mathbf{w}^*\|_2 = 1$ *and* $\gamma(\mathbf{w}^*) = \gamma > 0$*.*

4.2 Smooth Perceptron

In order to solve (3), the vanilla Perceptron repeatedly moves the direction of the classifier to that of examples on which it performs badly. However, this greedy policy only yields a sub-optimal convergence rate. To address this problem, Soheili & Pena (2012) propose the Smooth Perceptron, and the pseudo-code is summarized in the first box in Algorithm 1. The key idea is to find a classifier $\mathbf{v} \in \mathbb{R}^d$ that maximizes the following ℓ_2-regularized function:

$$\psi(\mathbf{v}) = -\tfrac{1}{2}\|\mathbf{v}\|_2^2 + \min_{\mathbf{q} \in \Delta^n} \mathbf{q}^\top A\mathbf{v}, \tag{4}$$

which is *non-smooth strongly-concave* with respect to $\mathbf{v}$. Under Assumption 1, Soheili & Pena (2012) show that the maximum value of $\psi(\mathbf{v})$ is $\max_{\mathbf{v} \in \mathbb{R}^d} \psi(\mathbf{v}) = \gamma^2/2$, and a classifier $\mathbf{v} \in \mathbb{R}^d$ satisfies (2) when $\psi(\mathbf{v}) \ge 0$. In order to solve (4), Soheili & Pena (2012) apply the classical Nesterov's excessive gap technique (Nesterov, 2005) for strongly concave functions. This algorithm introduces a smoothed approximation of (4), which is parameterized by some constant $\mu > 0$, and defined as

$$\psi_\mu(\mathbf{v}) = -\tfrac{1}{2}\|\mathbf{v}\|_2^2 + \min_{\mathbf{q} \in \Delta^n} \left[\mathbf{q}^\top A\mathbf{v} + \mu D_E\left(\mathbf{q}, \tfrac{\mathbf{1}}{n}\right)\right], \tag{5}$$

Algorithm 1 Smooth Perceptron (Soheili & Pena, 2012)

Initialization : $\theta_0 = \frac{2}{3}$, $\mu_0 = 4$, $\mathbf{v}_0 = \frac{A^\top \mathbf{1}}{n}$
Initialization : $\mathbf{q}_0 = \mathbf{q}_{\mu_0}(\mathbf{v}_0)$, where $\mathbf{q}_\mu(\mathbf{v}) = \operatorname{argmin}_{\mathbf{q}\in\Delta^n} \mathbf{q}^\top A\mathbf{v} + \mu D_E\left(\mathbf{q}, \frac{\mathbf{1}}{n}\right)$.
for $t = 1, \ldots, T-1$ **do**
  $\mathbf{v}_t = (1-\theta_{t-1})(\mathbf{v}_{t-1} + \theta_{t-1} A\mathbf{q}_{t-1}) + \theta_{t-1}^2 A\mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1})$
  $\mu_t = (1-\theta_{t-1})\mu_{t-1}$
  $\mathbf{q}_t = (1-\theta_{t-1})\mathbf{q}_{t-1} + \theta_{t-1}\mathbf{q}_{\mu_t}(\mathbf{v}_t)$
  $\theta_t = \frac{2}{t+3}$
end for
Output: $\mathbf{v}_{T-1}$

$$\mathsf{OL}^{\mathbf{w}} = \text{OFTL}\left[h_{t-1}(\cdot), \mathbb{R}^d\right] \Leftrightarrow \mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \sum_{j=1}^{t-1} \alpha_j h_j(\mathbf{w}) + \alpha_t h_{t-1}(\mathbf{w})$$

$$\mathsf{OL}^{\mathbf{p}} = \text{FTRL}^+\left[E(\cdot), \frac{\mathbf{1}}{n}, \frac{1}{4}, \Delta^n\right] \Leftrightarrow \mathbf{p}_t = \operatorname*{argmin}_{\mathbf{p}\in\Delta^n} \tfrac{1}{4}\textstyle\sum_{s=1}^t \alpha_s \ell_s(\mathbf{p}) + \mathcal{D}_E\left(\mathbf{p}, \frac{\mathbf{1}}{n}\right)$$

Output: $\overline{\mathbf{w}}_T$

which bounds the function in (4) from below, i.e., $\forall \mathbf{v} \in \mathbb{R}^d, \mu > 0$, $\psi_\mu(\mathbf{v}) \leq \psi(\mathbf{v}) + \mu \log n$. Then, the algorithm performs a sophisticated update rule such that the excessive gap condition holds $\forall t \geq 1$: $\gamma^2/2 \leq \|A\mathbf{v}_t\|^2/2 \leq \psi_\mu(\mathbf{v}_t)$. We refer to Soheili & Pena (2012) for more details.

Soheili & Pena (2012) show that the smooth Perceptron can output a positive-margin classifier after $\Omega(\sqrt{\log n}/\gamma)$ iterations. However, the analysis is quite involved, which heavily relies on the complicated relationship between $\psi(\mathbf{v}_t)$, $\psi_\mu(\mathbf{v}_t)$ and $\|A\mathbf{v}_t\|^2$. In the following, we provide a no-regret explanation of this algorithm and then show that the convergence rate can be easily obtained under our framework. Specifically, we define the objective function in Protocol 1 as

$$g(\mathbf{w}, \mathbf{p}) = \mathbf{p}^\top A\mathbf{w} - \tfrac{1}{2}\|\mathbf{w}\|_2^2, \tag{6}$$

and provide the equivalent expression in the second box of Algorithm 1. More specifically, we have the following proposition. The proof is given in Appendix C.1.

Proposition 1. *Let $\alpha_t = t$. Then the two interpretations of the smooth Perceptron in Algorithm 1 are the same, in the sense that $\mathbf{v}_{T-1} = \overline{\mathbf{w}}_T$, and $\mathbf{q}_{T-1} = \frac{\sum_{t=1}^T \alpha_t \mathbf{p}_t}{\sum_{t=1}^T \alpha_t}$.*

Proposition 1 shows that, under Protocol 1, the smooth Perceptron can be seen as optimizing (6) by implementing two online learning algorithms: in each round t, the $\mathbf{w}$-player applies OFTL with $\psi_t = h_{t-1}$, a decision which the $\mathbf{p}$-player subsequently observes and responds with FTRL$^+$. Moreover, we obtain theoretical guarantees for the smooth Perceptron based on Theorem 2, matching the convergence rate provided by Soheili & Pena (2012).

Theorem 3. *Let $\alpha_t = t$. Under Protocol 1, the regret of the two players of Algorithm 1 is bounded by $R^{\mathbf{w}} \leq 2\sum_{t=1}^T \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2$ and $R^{\mathbf{p}} \leq 4\log n - 2\sum_{t=1}^T \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2$. Moreover, $\overline{\mathbf{w}}_T$ has non-negative margin when $T = \Omega(\frac{\sqrt{\log n}}{\gamma})$.*

4.3 Accelerated Perceptron via ERM

Since binary linear classification is a (and perhaps the most fundamental) supervised machine learning problem, it can be naturally solved via empirical risk minimization (ERM). This idea is adopted by Ji et al. (2021), who consider the following ERM problem:

$$\min_{\mathbf{v}\in\mathbb{R}^d} R(\mathbf{v}) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \ell(-y^{(i)}\mathbf{v}^\top \mathbf{x}^{(i)}), \tag{7}$$

where $\ell(z) = \exp(z)$ is the exponential loss. To minimize (7), Ji et al. (2021) propose the a normalized momentum-based gradient descent (NMGD) algorithm, and show that NMGD can converge to

Algorithm 2 Accelerated Perceptron of Ji et al. (2021)

Input: $\mathbf{q}_0 = \frac{\mathbf{1}}{n}, \mathbf{v}_0 = \mathbf{0}, \mathbf{g}_0 = \mathbf{0}$.
for $t = 1, \ldots, T$ **do**
  Set $\theta_{t-1} = \frac{t}{2(t+1)}, \beta_t = \frac{t}{t+1}$
  $\mathbf{v}_t = \mathbf{v}_{t-1} - \theta_{t-1}(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1})$
  for $i = 1, \ldots, n$ **do**
    $q_{t,i} = \frac{\exp(-y^{(i)} \mathbf{v}_t^\top \mathbf{x}^{(i)})}{\sum_{j=1}^n \exp(-y^{(j)} \mathbf{v}_t^\top \mathbf{x}^{(j)})}$
  end for
  $\mathbf{g}_t = \beta_t(\mathbf{g}_{t-1} - A^\top \mathbf{q}_t)$
end for
Output: $\mathbf{v}_T$

$$\mathsf{OL}^{\mathbf{w}} = \text{OFTL}\left[h_{t-1}(\cdot), \mathbb{R}^d\right] \Leftrightarrow \mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^d} \sum_{j=1}^{t-1} \alpha_j h_j(\mathbf{w}) + \alpha_t h_{t-1}(\mathbf{w})$$

$$\mathsf{OL}^{\mathbf{p}} = \text{FTRL}^+\left[E(\cdot), \tfrac{\mathbf{1}}{n}, \tfrac{1}{4}, \Delta^n\right] \Leftrightarrow \mathbf{p}_t = \operatorname*{argmin}_{\mathbf{p} \in \Delta^n} \frac{1}{4} \sum_{s=1}^{t} \alpha_s \ell_s(\mathbf{p}) + \mathcal{D}_E\left(\mathbf{p}, \frac{\mathbf{1}}{n}\right)$$

Output: $\overline{\mathbf{w}}_T$

a classifier with a positive margin after $O(1/\gamma^2)$ iterations. Based on the property of the normalized gradient, they also provide primal-dual form of the algorithm (presented in the first-box of Algorithm 2), which can be considered as applying Nesterov's accelerated gradient descent in the *dual space* (Ji et al., 2021) .

For the accelerated Perceptron of Ji et al. (2021), we set $g(\mathbf{w}, \mathbf{p}) = \mathbf{p}^\top A\mathbf{w} - \frac{1}{2}\|\mathbf{w}\|_2^2$ and provide its equivalent form at the second box in Algorithm 2. Specifically, we have the following proposition.

Proposition 2. *Let $\alpha_t = t$. Then the two interpretations of the accelerated Perceptron in Algorithm 2 are the same, in the sense that $\mathbf{q}_T = \mathbf{p}_T$ and $\mathbf{v}_T = \frac{1}{4}\sum_{t=1}^T \alpha_t \mathbf{w}_t = \frac{1}{4}\left(\sum_{t=1}^T \alpha_t\right) \cdot \overline{\mathbf{w}}_T$.*

Remark Note that in Ji et al. (2021), the parameter θ_{t-1} is set to be 1. We observe that this causes a mismatch between the weights and $\mathbf{w}_t'$s in the output (when $\theta_{t-1} = 1$, $\mathbf{v}_T$ will become $\sum_{t=1}^T \alpha_{t+1}\mathbf{w}_t$ instead of $\sum_{t=1}^T \alpha_t \mathbf{w}_t$). Our no-regret analysis suggests that θ_{t-1} should be configured as $\frac{t}{2(t+1)}$, and later we will show that this procedure helps eliminate the $\log t$ factor in the convergence rate.

The proposition above reveals that the smooth Perceptron and the accelerated Perceptron of Ji et al. (2021) are closely related. The main difference is that the accelerated Perceptron of Ji et al. (2021) outputs the weighted sum of all $\mathbf{w}_t$'s, instead of the weighted average. The rationale behind this phenomenon is that, instead of the margin, Ji et al. (2021) use the *normalized* margin to measure the performance of the algorithm, defined as $\overline{\gamma}(\mathbf{v}) = \min_{\mathbf{p} \in \Delta^n} \mathbf{p}^\top A\mathbf{v}/\|\mathbf{v}\|_2$. They not only prove $\overline{\gamma}(\mathbf{v}_T) > 0$ (which directly implies $\gamma(\mathbf{v}_T) > 0$), but also show $\overline{\gamma}(\mathbf{v}_T) = \Omega(\gamma - \frac{\log T}{\gamma T^2})$. This guarantee is more powerful than the previous results, as it implies that the *normalized margin can be maximized* in an $O(\log t/t^2)$ rate. When t approaches ∞, the direction of $\mathbf{v}_T$ will converge to that of the maximal margin classifier. We show that a better margin-maximization rate can be directly obtained under our framework with the parameter setting in Algorithm 2.

Theorem 4. *Let $\alpha_t = t$. Under Protocol 1, the regret of the two players of Algorithm 2 is bounded by $R^{\mathbf{w}} \le 2\sum_{t=1}^T \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2$ and $R^{\mathbf{p}} \le 4\log n - 2\sum_{t=1}^T \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2$. Moreover, $\overline{\mathbf{w}}_T$ has a non-negative margin when $T = \Omega(\sqrt{\log n}/\gamma)$ and $\overline{\gamma}(\mathbf{v}_T) = \Omega(\gamma - \frac{8\log n}{\gamma T(T+1)})$.*

Algorithm 3 NAG

Input: $\mathbf{v}_0 = \mathbf{0}, \mathbf{s}_0 = \mathbf{0}$.
for $t = 1, \dots, T$ **do**
  $\mathbf{u}_t = \mathbf{s}_{t-1} + \frac{1}{2(t-1)}\mathbf{v}_{t-1}$
  $\mathbf{v}_t = \mathbf{v}_{t-1} - \eta_t \nabla R(\mathbf{u}_t)$
  $\mathbf{s}_t = \mathbf{s}_{t-1} + \frac{1}{2(t+1)}\mathbf{v}_t$
end for
Output: $\mathbf{s}_T$

$$\mathsf{OL}^{\mathbf{p}} = \text{OFTRL}\left[E(\cdot), \tfrac{\mathbf{1}}{n}, \tfrac{1}{4}, \ell_{t-1}(\cdot), \Delta^n\right] \Leftrightarrow \mathbf{p}_t = \operatorname*{argmin}_{\mathbf{p}\in\Delta^n} \frac{1}{4}\left[\sum_{s=1}^{t-1} \alpha_s \ell_s(\mathbf{p}) + \alpha_t \ell_{t-1}(\mathbf{p})\right] + \mathcal{D}_E\left(\mathbf{p}, \tfrac{\mathbf{1}}{n}\right)$$

$$\mathsf{OL}^{\mathbf{w}} = \text{FTRL}^{+}[0, 0, 1, \mathbb{R}^d] \Leftrightarrow \mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \sum_{j=1}^{t} \alpha_j h_j(\mathbf{w})$$

Output: $\widetilde{\mathbf{w}}_T = \frac{\sum_{t=1}^T \alpha_t}{4}\overline{\mathbf{w}}_T$.

4.4 Accelerated Perceptron of Yu et al. (2014)

Finally, Yu et al. (2014) considers applying the mirror-prox algorithm to solve the max-min optimization problem:

$$\max_{\|\mathbf{w}\|_2\le 1} \min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{w}, \tag{8}$$

As discovered by Rakhlin & Sridharan (2013), mirror-prox can be recovered by their optimistic online mirror descent (OMD) with an appropriately chosen optimistic term. Here, we show that, by simply manipulating the notation, the algorithm of Yu et al. (2014) can be recovered as two-players applying OMD to solve (8). Due to page limitation, the algorithmic details and their analysis are postponed to Appendix C.5, which we summarize as follows (formalized in Theorems 8 and 9):

Theorem 5 (informal)**.** *Let $g(\mathbf{w}, \mathbf{p}) = \mathbf{p}^\top A\mathbf{w}$. The Perceptron of Yu et al. (2014) can be described under Protocol 1, where both players apply OMD endowed with the appropriate parameters. Moreover, under Assumption 1, $\overline{\mathbf{w}}_T$ has a non-negative margin when $T = \Omega\left(\frac{\sqrt{\log n}}{\gamma}\right)$.*

5 Beyond Perceptrons

In This section, we show that our framework benefits a wide range of other problems.

5.1 Margin Maximization by Nesterov's Accelerated Gradient Descent

One major task for the implicit bias study is to explain why *commonly-used* first-order optimization methods generalize well. For linearly separable data, previous work proves that GD and momentum-based GD prefers the ℓ_2-maximal margin classifier by showing that they can maximize the margin. In this part, we show that the well-known Nesterov's accelerated gradient descent (NAG), with appropriately chosen parameters, can maximize the margin in an $O(1/\gamma^2)$ rate. Specifically, following previous work, we consider maximize the margin by solving the ERM problem in (7) and apply Nesterov's accelerated gradient descent (NAG) to minimize the objective function. The details of the algorithm are summarized in the first box of Algorithm 3, and its equivalent form under Protocol 1 is presented in the second box of Algorithm 3. For the NAG algorithm, we set the objective function as $g(\mathbf{w}, \mathbf{p}) = \mathbf{p}^\top A\mathbf{w} - \frac{1}{2}\|\mathbf{w}\|_2^2$, and have the following conclusion.

Theorem 6. *Let $\eta_t = \frac{t}{R(\mathbf{u}_t)}$ and $\alpha_t = t$. Then, the two expressions in Algorithm 3 are equivalent, in the sense that $\mathbf{s}_T = \widetilde{\mathbf{w}}_T$. Moreover, $\frac{\min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{s}_T}{\|\mathbf{s}_T\|_2} \ge \gamma - \frac{8\log n + 2}{T(T+1)\gamma}$.*

Algorithm 4 Accelerated algorithm for the p-norm perceptron

$$\mathsf{OL}^{\mathbf{w}} = \text{OFTRL}\left[\tfrac{1}{2(q-1)}\|\cdot\|_q^2, \mathbf{0}, \eta^{\mathbf{w}}, h_{t-1}(\cdot), \mathbb{R}^d\right] \Leftrightarrow$$

$$\mathbf{w}_t = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \eta^{\mathbf{w}} \sum_{j=1}^{t-1} \alpha_j h_j(\mathbf{w}) + \alpha_t h_{t-1}(\mathbf{w}) + D_{\frac{1}{2(q-1)}\|\cdot\|_q^2}(\mathbf{w}, \mathbf{0})$$

$$\mathsf{OL}^{\mathbf{p}} = \text{FTRL}^+\left[E, \frac{\mathbf{1}}{n}, \eta^p, \Delta^n\right] \Leftrightarrow \mathbf{p}_t = \operatorname*{argmin}_{\mathbf{p}\in\Delta^n} \eta^{\mathbf{p}} \sum_{s=1}^{t} \alpha_s \ell_s(\mathbf{p}) + \mathcal{D}_E\left(\mathbf{p}, \frac{\mathbf{1}}{n}\right)$$

Output: $\overline{\mathbf{w}}_T = \frac{1}{T}\sum_{s=1}^{T} \mathbf{w}_s$.

Remark Theorem 6 indicates that NAG can also maximize the margin in an $O(1/T^2)$ rate. Note that, *in Algorithm 3, the* $\mathbf{p}$*-player plays first and the* $\mathbf{w}$*-player second, while, in Algorithm 2, it is the other way around*. This difference makes sense as Algorithm 3 can be considered as applying Nesterov's acceleration in the dual space (Ji et al., 2021), while Algorithm 3 uses Nesterov's acceleration in the *primal space*.

5.2 Accelerated p-norm Perceptron

In this section, we focus on the p-norm Perceptron problem, introduced by Gentile (2000). Compared to the classical perceptron problem, p-norm Perceptron introduces the following assumption, which is more general than Assumption 1.

Assumption 2. *For the p-norm Perceptron problem, we assume: $\forall i \in [n]$, $\|\mathbf{x}^{(i)}\|_p \le 1$, where $p \in [2, \infty)$. Moreover, assume there exists a $\mathbf{w}^* \in \mathbb{R}^d$, such that $\|\mathbf{w}^*\|_q \le 1$ and $\min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{w}^* \ge \gamma$. Here, $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$; i.e., $\frac{1}{p}+\frac{1}{q}=1$ and $q \in (1, 2]$.*

Under Assumption 2, Gentile (2000) proposes a mirror-descent style algorithm that achieves an $\Omega((1-p)/\gamma^2)$ convergence rate. In the following, we provide a new algorithm with a better rate. Specifically, under Protocol 1, we define the objective function as $g(\mathbf{w}, \mathbf{p}) = \mathbf{p}^\top A\mathbf{w}$, and introduce Algorithm 4, wherein the $\mathbf{w}$-player uses OFTRL with regularizer $\frac{1}{2(q-1)}\|\cdot\|_q^2$, which is 1-strongly convex w.r.t. the q-norm (Orabona, 2019). On the other hand, the $\mathbf{q}$-player employs the FTRL$^+$ algorithm. We have the following result.

Theorem 7. *Let $\alpha_t = 1$, $\eta^{\mathbf{p}} = 1/\eta^{\mathbf{w}}$, and $\eta^{\mathbf{w}} = \sqrt{\frac{1}{2(q-1)\log n}}$. Then the output $\overline{\mathbf{w}}_T$ of Algorithm 4 has a non-negative margin when $T = \Omega\left(\sqrt{2(p-1)\log n}/\gamma\right)$.*

6 Conclusion

In this paper, we provide a unified analysis for the existing accelerated Perceptrons, and obtain improved results for a series of problems. In the future, we will explore how to extend our framework to other closely related areas, such as semi-definite programming (Garber & Hazan, 2011) and generalized margin maximization (Sun et al., 2022). Another interesting direction is to consider whether other more advanced online learning algorithms (Orabona & Pál, 2016; Cutkosky & Orabona, 2018; Jun et al., 2017; Zhang et al., 2018; Wang et al., 2020; Zhao et al., 2020) is useful for perceptron and the implicit bias analysis based on our framework.

Acknowledgments

We gratefully thank the AI4OPT Institute for funding, as part of NSF Award 2112533. We also acknowledge the NSF for their support through Award IIS-1910077, as well as an ARC-ACO fellowship provided by Georgia Tech.

REFERENCES

Jacob Abernethy, Kevin A Lai, Kfir Y Levy, and Jun-Kun Wang. Faster rates for convex-concave games. In *Proceedings of the 31st Annual Conference on Learning Theory*, pp. 1595–1625, 2018.

Jacob D Abernethy and Jun-Kun Wang. On frank-wolfe and equilibrium computation. *Advances in Neural Information Processing Systems 30*, 30, 2017.

Mark A Aizerman. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. A second-order perceptron algorithm. *SIAM Journal on Computing*, 34(3):640–668, 2005.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pp. 1493–1529, 2018.

Daniel Dadush, Laszlo A Vegh, and Giacomo Zambelli. Rescaling algorithms for linear conic feasibility. *Mathematics of Operations Research*, 45(2):732–754, 2020.

C Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *10th Innovations in Theoretical Computer Science conference*, 2019.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations*, 2018.

John Dunagan and Santosh Vempala. A polynomial-time rescaling algorithm for solving linear programs. In *Proceedings of the 36th annual ACM symposium on Theory of computing*, pp. 28, 2004.

Yoav Freund and Robert E Schapire. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296, 1999.

Dan Garber and Elad Hazan. Approximating semidefinite programs in sublinear time. *Advances in Neural Information Processing Systems 24*, 2011.

Claudio Gentile. A new approximate maximal margin classification algorithm. *Advances in Neural Information Processing Systems 13*, 2000.

Claudio Gentile. A new approximate maximal margin classification algorithm. *Journal of Machine Learning Research*, 2:213–242, 2001.

Claudio Gentile and Nick Littlestone. The robustness of the p-norm algorithms. In *Proceedings of the 12th annual conference on Computational learning theory*, pp. 1–11, 1999.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 772–804, 2021.

Ziwei Ji, Nathan Srebro, and Matus Telgarsky. Fast margin maximization via dual acceleration. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4860–4869, 2021.

Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved strongly adaptive online learning using coin betting. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 943–951, 2017.

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

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428, 2019.

Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yu Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249, 2005.

Yurii Nesterov. On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonomika i Mateaticheskie Metody*, 24(3):509–517, 1988.

A.B.J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pp. 615–622, 1962.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems 29*, 2016.

Javier Peña and Negar Soheili. A deterministic rescaled perceptron algorithm. *Mathematical Programming*, 155(1):497–510, 2016.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems 26*, pp. 3066–3074, 2013.

Aaditya Ramdas and Javier Pena. Towards a deeper geometric, analytic and algorithmic understanding of margins. *Optimization Methods and Software*, 31(2):377–391, 2016.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

Negar Soheili and Javier Pena. A smooth perceptron algorithm. *SIAM Journal on Optimization*, 22 (2):728–737, 2012.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19 (1):2822–2878, 2018.

Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. *arXiv preprint arXiv:2205.12808*, 2022.

Guanghui Wang, Shiyin Lu, Quan Cheng, Wei-wei Tu, and Lijun Zhang. Sadam: A variant of adam for strongly convex functions. In *International Conference on Learning Representations*, 2020.

Jun-Kun Wang and Jacob D Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems 32*, pp. 3824–3834, 2018.

Jun-Kun Wang, Jacob Abernethy, and Kfir Y Levy. No-regret dynamics in the fenchel game: A unified framework for algorithmic convex optimization. *arXiv preprint arXiv:2111.11309*, 2021.

A Wendemuth. Learning the unlearnable. *Journal of Physics A: Mathematical and General*, 28(18): 5423, 1995.

Adams Wei Yu, Fatma Kilinc-Karzan, and Jaime Carbonell. Saddle points and accelerated perceptron algorithms. In *International Conference on Machine Learning*, pp. 1827–1835, 2014.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems 31*, pp. 1323–1333, 2018.

Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. *Advances in Neural Information Processing Systems 33*, pp. 12510–12520, 2020.

Protocol 2 A set of useful algorithms for the weighted OCO

for $t = 1, \ldots, T$ **do**

$$
\begin{aligned}
\mathrm{OFTRL}[R, \mathbf{z}_0, \eta, \psi_t, \mathcal{Z}] : \mathbf{z}_t &= \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \eta \left[\sum_{j=1}^{t-1} \alpha_j f_j(\mathbf{z}) + \alpha_t \psi_t(\mathbf{z})\right] + D_R(\mathbf{z}, \mathbf{z}_0) \\
\mathrm{OMD}[R, \widehat{\mathbf{z}}_0, \eta, \psi_t, \mathcal{Z}] : \mathbf{z}_t &= \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \eta \left\langle \alpha_t \nabla \psi_t(\mathbf{z}_{t-1}), \mathbf{z}\right\rangle + \mathcal{D}_R(\mathbf{z}, \widehat{\mathbf{z}}_{t-1}) \\
\widehat{\mathbf{z}}_t &= \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \eta \left\langle \alpha_t \nabla f_t(\mathbf{z}_t), \mathbf{z}\right\rangle + \mathcal{D}_R(\mathbf{z}, \widehat{\mathbf{z}}_{t-1}) \\
\mathrm{OFTL}[\psi_t, \mathcal{Z}] : \mathbf{z}_t &= \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \sum_{j=1}^{t-1} \alpha_j f_j(\mathbf{z}) + \alpha_t \psi_t(\mathbf{z}) \\
\mathrm{FTRL}^+[R, \mathbf{z}_0, \eta, \mathcal{Z}] : \mathbf{z}_t &= \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \eta \cdot \sum_{j=1}^{t} \alpha_j f_j(\mathbf{z}) + \mathcal{D}_R(\mathbf{z}, \mathbf{z}_0)
\end{aligned}
$$

end for

A Regret bounds on OCO algorithms

Lemma 1 (Lemma 13 of Wang et al. (2021)). *Let $\widetilde{\mathbf{z}}_t = \operatorname{argmin}_{\mathbf{z}\in\mathcal{Z}} \sum_{s=1}^{t-1} \alpha_s f_s(\mathbf{z}_s)$. Assume f_t and ψ_t are λ-strongly convex w.r.t. some norm $\|\cdot\|$. Then, for the OFTL$[\psi_t, \mathcal{Z}]$ algorithm, the regret is bounded by*

$$
\begin{aligned}
\forall \mathbf{z} \in \mathcal{Z},\ \sum_{t=1}^{T} \alpha_t f_t(\mathbf{z}_t) - \sum_{t=1}^{T} \alpha_t f_t(\mathbf{z}) \le \sum_{t=1}^{T} \alpha_t (f_t(\mathbf{z}_t) - f_t(\widetilde{\mathbf{z}}_{t+1})) - \alpha_t(\psi_t(\mathbf{z}_t) - \psi_t(\widetilde{\mathbf{z}}_{t+1})) \\
- \frac{1}{2} \sum_{t=1}^{T} \lambda \alpha_t \|\mathbf{z}_t - \widetilde{\mathbf{z}}_{t+1}\|^2.
\end{aligned} \tag{9}
$$

Lemma 2 (Lemma 13 of Wang et al. (2021)). *For FTRL$^+[R, \mathbf{z}_0, \eta, \mathcal{Z}]$ algorithm, suppose R is β-strongly convex w.r.t some norm $\|\cdot\|$ and each $f_t(\cdot)$ is λ-strongly convex w.r.t. the same norm, where λ may be 0. Then*

$$
\forall \mathbf{z} \in \mathcal{Z},\ \sum_{t=1}^{T} \alpha_t f_t(\mathbf{z}_t) - \sum_{t=1}^{T} \alpha_t f_t(\mathbf{z}) \le \frac{R(\mathbf{z}) - R(\mathbf{z}_0)}{\eta} - \sum_{t=1}^{T} \left(\frac{\lambda \sum_{s=1}^{t-1} \alpha_s}{2} + \frac{\beta}{2\eta} \right) \|\mathbf{z}_t - \mathbf{z}_{t-1}\|^2. \tag{10}
$$

Lemma 3 (Theorem 7.35 of Orabona (2019)). *For the OFTRL$[R, \mathbf{z}_0, \eta, \psi_t, \mathcal{Z}]$ algorithm, assume R is 1-strongly convex w.r.t. some norm $\|\cdot\|$. Then the regret is bounded by*

$$
\forall \mathbf{z} \in \mathcal{Z},\ \sum_{t=1}^{T} \alpha_t f_t(\mathbf{z}_t) - \sum_{t=1}^{T} \alpha_t f_t(\mathbf{z}) \le \frac{R(\mathbf{z}) - R(\mathbf{z}_0)}{\eta} + \sum_{t=1}^{T} \left[\frac{\|\alpha_t \nabla f_t(\mathbf{z}_t) - \alpha_t \nabla \psi_t(\mathbf{z}_t)\|_*^2}{2/\eta}. \right] \tag{11}
$$

Lemma 4 (Lemma 1 of Rakhlin & Sridharan (2013)). *For the OMD$[R, \widehat{\mathbf{z}}_0, \eta, \psi_t, \mathcal{Z}]$ algorithm, suppose R is 1-strongly convex w.r.t. some norm $\|\cdot\|$. Then the regret is bounded by*

$$
\begin{aligned}
\forall \mathbf{z} \in \mathcal{Z},\ \sum_{t=1}^{T} \alpha_t f_t(\mathbf{z}_t) - \sum_{t=1}^{T} \alpha_t f_t(\mathbf{z}) \le \frac{\mathcal{D}_R(\mathbf{z}, \widehat{\mathbf{z}}_0)}{\eta} + \sum_{t=1}^{T} \|\alpha_t \nabla f_t(\mathbf{z}_t) - \alpha_t \nabla \psi_t(\hat{\mathbf{z}}_{t-1})\|_* \|\mathbf{z}_t - \widehat{\mathbf{z}}_t\| \\
- \frac{1}{2\eta} \sum_{t=1}^{T} \left(\|\mathbf{z}_t - \widehat{\mathbf{z}}_t\|^2 + \|\mathbf{z}_t - \widehat{\mathbf{z}}_{t-1}\|^2 \right).
\end{aligned} \tag{12}
$$

B PROOF OF THEOREM 2

The proof is motivated by Abernethy et al. (2018). Let $\overline{R}^{\mathbf{w}}$ and $\overline{R}^{\mathbf{p}}$ denote the average regret of the players. On one hand, we have

$$\begin{aligned}
\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t g(\mathbf{w}_t,\mathbf{p}_t) &= \frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}-\alpha_t h_t(\mathbf{w}_t)\\
&= -\min_{\mathbf{w}'\in\mathcal{W}}\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t h_t(\mathbf{w}') - \overline{R}^{\mathbf{w}}\\
&= \max_{\mathbf{w}'\in\mathcal{W}}\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t g(\mathbf{w}',\mathbf{p}_t) - \overline{R}^{\mathbf{w}}\\
&\geq \frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t g(\mathbf{w},\mathbf{p}_t) - \overline{R}^{\mathbf{w}}\\
&\geq g(\mathbf{w},\overline{\mathbf{p}}_T) - \overline{R}^{\mathbf{w}}\\
&\geq \min_{\mathbf{p}\in\mathcal{Q}} g(\mathbf{w},\mathbf{p}) - \overline{R}^{\mathbf{w}}.
\end{aligned}$$

where $\mathbf{w}\in\mathcal{W}$ is an arbitrary point. On the other hand, we have

$$\begin{aligned}
\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t g(\mathbf{w}_t,\mathbf{p}_t) &= \frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t \ell_t(\mathbf{p}_t)\\
&= \min_{\mathbf{p}\in\mathcal{Q}}\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t \ell_t(\mathbf{p}) + \overline{R}^{\mathbf{p}}\\
&= \min_{\mathbf{p}\in\mathcal{Q}}\frac{1}{\sum_{t=1}^{T}\alpha_t}\sum_{t=1}^{T}\alpha_t g(\mathbf{w}_t,\mathbf{p}) + \overline{R}^{\mathbf{p}}\\
&\leq \min_{\mathbf{p}\in\mathcal{Q}} g(\overline{\mathbf{w}}_T,\mathbf{p}) + \overline{R}^{\mathbf{p}}.
\end{aligned}$$

C PROOF OF SECTION 4

In this section, we provide the detailed proofs of Section 4.

C.1 PROOF OF PROPOSITION 1

Recall the update rule of $\mathbf{w}$ and $\mathbf{p}$ in Algorithm 1:

$$\begin{aligned}
\mathbf{w}_t &= \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d}\sum_{j=1}^{t-1}\alpha_j h_j(\mathbf{w}) + \alpha_t h_{t-1}(\mathbf{w})\\
&= \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d}\sum_{j=1}^{t-1}-\alpha_j\mathbf{p}_j^{\top}A\mathbf{w} - \alpha_t\mathbf{p}_{t-1}^{\top}A\mathbf{w} + \frac{\sum_{s=1}^{t}\alpha_s}{2}\|\mathbf{w}\|_2^2\\
&= \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d}\sum_{j=1}^{t-1}-\alpha_j\mathbf{p}_j^{\top}A\mathbf{w} - \alpha_t\mathbf{p}_{t-1}^{\top}A\mathbf{w} + \frac{t(t+1)}{4}\|\mathbf{w}\|_2^2,
\end{aligned}$$

and

$$\mathbf{p}_t = \operatorname*{argmin}_{\mathbf{p}\in\Delta^n}\frac{1}{4}\sum_{s=1}^{t}\alpha_s\ell_s(\mathbf{p}) + \mathcal{D}_E\left(\mathbf{p},\frac{\mathbf{1}}{n}\right) = \operatorname*{argmin}_{\mathbf{p}\in\Delta^n}\sum_{s=1}^{t}\alpha_s\mathbf{p}^{\top}A\mathbf{w}_s + 4\mathcal{D}_E\left(\mathbf{p},\frac{\mathbf{1}}{n}\right), \quad (13)$$

Note that here we drop the $\|\mathbf{w}_s\|^2/2$ of ℓ_s in (13), as it is constant with respect to $\mathbf{p}$. We start from the initialization: first, from the fact that $\mathbf{p}_0 = \frac{\mathbf{1}}{n}$, we obtain

$$\overline{\mathbf{w}}_1 = \mathbf{w}_1 = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} h_0(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} -\mathbf{p}_0^\top A\mathbf{w} + \frac{1}{2}\|\mathbf{w}\|_2^2 = \frac{A^\top \mathbf{1}}{n} = \mathbf{v}_0.$$

Additionally,

$$\begin{aligned}
\mathbf{p}_1 =& \operatorname*{argmin}_{\mathbf{p}\in\Delta^n} \ell_1(\mathbf{p}) + 4\mathcal{D}_E\left(\mathbf{p}, \frac{\mathbf{1}}{n}\right) \\
=& \operatorname*{argmin}_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{w}_1 + 4\mathcal{D}_E\left(\mathbf{p}, \frac{\mathbf{1}}{n}\right) \\
=& \operatorname*{argmin}_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{v}_0 + 4\mathcal{D}_E\left(\mathbf{p}, \frac{\mathbf{1}}{n}\right) = \mathbf{q}_{\mu_0}(\mathbf{v}_0) = \mathbf{q}_0.
\end{aligned}$$

Next, we focus on $\mathbf{v}_t$ and its connection to $\overline{\mathbf{w}}_t$. We have

$$\begin{aligned}
\mathbf{v}_t =& (1-\theta_{t-1})(\mathbf{v}_{t-1} + \theta_{t-1}A^\top \mathbf{q}_{t-1}) + \theta_{t-1}^2 A^\top \mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1}) \\
=& (1-\theta_{t-1})\mathbf{v}_{t-1} + \theta_{t-1}A^\top\left((1-\theta_{t-1})\mathbf{q}_{t-1} + \theta_{t-1}\mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1})\right).
\end{aligned} \tag{14}$$

Define $\mathbf{r}_0 = \mathbf{v}_0$, and $\mathbf{r}_t = (1-\theta_{t-1})A^\top \mathbf{q}_{t-1} + \theta_{t-1}A^\top \mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1})$ for $t \geq 1$. We will show that

$$\mathbf{v}_t = \frac{\sum_{s=0}^t \alpha_{s+1}\mathbf{r}_s}{\sum_{s=0}^t \alpha_{s+1}}$$

for all $t \geq 0$ by induction. First, by definition, we have $\mathbf{v}_0 = \mathbf{r}_0 = \frac{\alpha_1 \mathbf{r}_0}{\alpha_1}$. Next, assume

$$\mathbf{v}_{t-1} = \frac{\sum_{s=0}^{t-1} \alpha_{s+1}\mathbf{r}_s}{\sum_{s=0}^{t-1} \alpha_{s+1}}.$$

Then

$$\begin{aligned}
\mathbf{v}_t &= (1-\theta_{t-1})\mathbf{v}_{t-1} + \theta_{t-1}\mathbf{r}_t \\
&= \frac{t}{t+2}\mathbf{v}_{t-1} + \frac{2}{t+2}\mathbf{r}_t \\
&= \frac{t}{t+2}\frac{1}{\sum_{s=0}^{t-1}\alpha_{s+1}}\sum_{s=0}^{t-1}\alpha_{s+1}\mathbf{r}_s + \frac{2}{(t+2)(t+1)}(t+1)\mathbf{r}_t \\
&= \frac{t}{t+2}\frac{2}{(t+1)t}\sum_{s=0}^{t-1}\alpha_{s+1}\mathbf{r}_s + \frac{2}{(t+2)(t+1)}\alpha_{t+1}\mathbf{r}_t \\
&= \frac{2}{(t+1)(t+2)}\sum_{s=0}^{t}\alpha_{s+1}\mathbf{r}_s = \frac{\sum_{s=0}^t \alpha_{s+1}\mathbf{r}_s}{\sum_{s=0}^t \alpha_{s+1}}.
\end{aligned} \tag{15}$$

Following a similar procedure, we obtain

$$\mathbf{q}_t = \frac{\sum_{s=0}^t \alpha_{s+1}\mathbf{q}_{\mu_s}(\mathbf{v}_s)}{\sum_{s=0}^t \alpha_{s+1}}.$$

Thus, for $t \geq 1$, we have

$$\begin{aligned}
\mathbf{r}_t &= (1-\theta_{t-1})A^\top \mathbf{q}_{t-1} + \theta_{t-1}A^\top \mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1}) \\
&= \frac{t}{t+2}\frac{2}{t(t+1)}A^\top\left(\sum_{s=0}^{t-1}\alpha_{s+1}\mathbf{q}_{\mu_s}(\mathbf{v}_s)\right) + \frac{2}{(t+2)(t+1)}(t+1)A^\top \mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1}) \\
&= \frac{2}{(t+1)(t+2)}A^\top\left(\sum_{s=0}^{t-1}\alpha_{s+1}\mathbf{q}_{\mu_s}(\mathbf{v}_s) + \alpha_{t+1}\mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1})\right) \\
&= \operatorname*{argmin}_{\mathbf{r}\in\mathbb{R}^d} -\left(\sum_{s=0}^{t-1}\alpha_{s+1}\mathbf{q}_{\mu_s}(\mathbf{v}_s) + \alpha_{t+1}\mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1})\right)^\top A\mathbf{r} + \frac{(t+1)(t+2)}{4}\|\mathbf{r}\|_2^2.
\end{aligned} \tag{16}$$

Finally, note that

$$\mu_t = 4\prod_{i=1}^{t}\frac{i}{i+2} = 4\frac{1}{1+2}\cdot\ldots\frac{t}{t+2} = 4\frac{2}{(t+1)(t+2)} = \frac{4}{\sum_{s=0}^{t}\alpha_{s+1}},$$

so

$$\begin{aligned}\mathbf{q}_{\mu_t}(\mathbf{v}_t) &= \operatorname*{argmin}_{\mathbf{q}\in\Delta^n}\mathbf{q}^\top A\left(\frac{\sum_{s=0}^{t}\alpha_{s+1}\mathbf{r}_s}{\sum_{s=0}^{t}\alpha_{s+1}}\right) + \frac{4}{\sum_{s=0}^{t}\alpha_{s+1}}\mathcal{D}_E\left(\mathbf{q},\frac{\mathbf{1}}{n}\right)\\ &= \operatorname*{argmin}_{\mathbf{q}\in\Delta^n}\mathbf{q}^\top A\left(\sum_{s=0}^{t}\alpha_{s+1}\mathbf{r}_s\right) + 4\mathcal{D}_E\left(\mathbf{q},\frac{\mathbf{1}}{n}\right).\end{aligned} \tag{17}$$

To summarize, for $t=1$, we have $\mathbf{r}_0 = \mathbf{v}_0 = \mathbf{w}_1$, and $\mathbf{q}_{\mu_0}(\mathbf{v}_0) = \mathbf{p}_1$. For $t \geq 2$, we know that

$$\mathbf{r}_{t-1} = \operatorname*{argmax}_{\mathbf{r}\in\mathbb{R}^d} -\left(\sum_{s=0}^{t-2}\alpha_{s+1}\mathbf{q}_{\mu_s}(\mathbf{v}_s) + \alpha_t\mathbf{q}_{\mu_{t-2}}(\mathbf{v}_{t-2})\right)^\top A\mathbf{r} + \frac{t(t+1)}{4}\|\mathbf{r}\|_2^2, \tag{18}$$

and

$$\mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1}) = \operatorname*{argmin}_{\mathbf{q}\in\Delta^n}\mathbf{q}^\top A\left(\sum_{s=0}^{t-1}\alpha_{s+1}\mathbf{r}_s\right) + 4\mathcal{D}_E\left(\mathbf{q},\frac{\mathbf{1}}{n}\right).$$

The proof is finished by replacing $\mathbf{q}_{\mu_{t-1}}(\mathbf{v}_{t-1})$ as $\mathbf{p}_t$, $\mathbf{r}_{t-1}$ as $\mathbf{w}_t$, and noticing that

$$\mathbf{v}_{T-1} = \frac{1}{\sum_{s=0}^{T-1}\alpha_{s+1}}\sum_{s=0}^{T-1}\alpha_{s+1}\mathbf{r}_s = \frac{1}{\sum_{s=1}^{T}\alpha_s}\sum_{s=1}^{T}\alpha_s\mathbf{w}_s = \overline{\mathbf{w}}_T.$$

and

$$\mathbf{q}_{T-1} = \frac{1}{\sum_{s=0}^{T-1}\alpha_{s+1}}\sum_{s=0}^{T-1}\alpha_{s+1}\mathbf{q}_{\mu_s}(\mathbf{v}_s) = \frac{1}{\sum_{s=1}^{T}\alpha_s}\sum_{s=1}^{T}\alpha_s\mathbf{p}_s$$

C.2 Proof of Theorem 3

Let $\widetilde{\mathbf{w}}_t = \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d}\sum_{s=1}^{t-1}\alpha_s h_s(\mathbf{w})$. Note that $h_t(\cdot)$ is 1-strongly convex w.r.t $\|\cdot\|_2$. Thus, for the $\mathbf{w}$-player, based on Lemma 1, we have

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t h_t(\mathbf{w}_t) - \alpha_t h_t(\mathbf{w}) &\leq \sum_{t=1}^{T}\alpha_t\left(h_t(\mathbf{w}_t) - h_t(\widetilde{\mathbf{w}}_{t+1})\right) - \alpha_t\left(h_{t-1}(\mathbf{w}_t) - h_{t-1}(\widetilde{\mathbf{w}}_{t+1})\right)\\ &= \sum_{t=1}^{T}\alpha_t\left(-\mathbf{p}_t^\top A\mathbf{w}_t + \frac{1}{2}\|\mathbf{w}_t\|_2^2 + \mathbf{p}_t^\top A\widetilde{\mathbf{w}}_{t+1} - \frac{1}{2}\|\widetilde{\mathbf{w}}_{t+1}\|_2^2\right)\\ &\quad - \alpha_t\left(-\mathbf{p}_{t-1}^\top A\mathbf{w}_t + \frac{1}{2}\|\mathbf{w}_t\|_2^2 + \mathbf{p}_{t-1}^\top A\widetilde{\mathbf{w}}_{t+1} - \frac{1}{2}\|\widetilde{\mathbf{w}}_{t+1}\|_2^2\right)\\ &= \sum_{t=1}^{T}\alpha_t\left(\mathbf{p}_t - \mathbf{p}_{t-1}\right)^\top A(\widetilde{\mathbf{w}}_{t+1} - \mathbf{w}_t).\end{aligned} \tag{19}$$

Next, according to the definition of $\mathbf{w}_t$ and $\widetilde{\mathbf{w}}_{t+1}$, we have

$$\mathbf{w}_t = \frac{1}{\sum_{s=1}^{t}\alpha_s}A^\top\left(\sum_{s=1}^{t-1}\alpha_s\mathbf{p}_s + \alpha_t\mathbf{p}_{t-1}\right),$$

and

$$\widetilde{\mathbf{w}}_{t+1} = \frac{1}{\sum_{s=1}^{t}\alpha_s}A^\top\left(\sum_{s=1}^{t-1}\alpha_s\mathbf{p}_s + \alpha_t\mathbf{p}_t\right).$$

Combining the two equations above with (19), we have that for any $\mathbf{w} \in \mathbb{R}^d$,

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t h_t(\mathbf{w}_t) - \alpha_t h_t(\mathbf{w}) &\leq \sum_{t=1}^{T} \frac{\alpha_t^2}{\sum_{s=1}^{t} \alpha_s} \|A^\top(\mathbf{p}_t - \mathbf{p}_{t-1})\|_2^2 \\
&= \sum_{t=1}^{T} \frac{\alpha_t^2}{\sum_{s=1}^{t} \alpha_s} \left\| \sum_{i=1}^{n} y^{(i)} [\mathbf{p}_t - \mathbf{p}_{t-1}]_i \mathbf{x}^{(i)} \right\|_2^2 \\
&\overset{(2)}{\leq} \sum_{t=1}^{T} \frac{\alpha_t^2}{\sum_{s=1}^{t} \alpha_s} \left(\sum_{i=1}^{n} |\mathbf{p}_t - \mathbf{p}_{t-1}|_i \|\mathbf{x}^{(i)}\|_2 \right)^2 \\
&\overset{(3)}{\leq} \sum_{t=1}^{T} \frac{\alpha_t^2}{\sum_{s=1}^{t} \alpha_s} \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2 \\
&= \sum_{t=1}^{T} \frac{2t^2}{t(t+1)} \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2 \overset{(4)}{\leq} 2 \sum_{t=1}^{T} \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2,
\end{aligned} \tag{20}$$

where inequality (2) is based on the triangle inequality, inequality (3) is derived from Assumption 1, and inequality (4) is because $\frac{t}{t+1} \leq 1$.

On the other hand, for the $\mathbf{p}$-player, note that the the regularizer is 1-strongly convex w.r.t. the $\|\cdot\|_1$. Thus, according to Lemma 2 (with $\lambda = 0, \beta = 1$), we have

$$\sum_{t=1}^{T} \alpha_t \ell_t(\mathbf{p}_t) - \sum_{t=1}^{T} \alpha_t \ell_t(\mathbf{p}) \leq 4 \log n - 2 \sum_{t=1}^{T} \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2. \tag{21}$$

for any $\mathbf{p} \in \Delta^n$. Finally, we focus on the iteration complexity. First, define $\widehat{\mathbf{w}}^* = \gamma \mathbf{w}^*$, where $\mathbf{w}^*$ is the maximum margin classifier defined in Assumption 1. Then we have

$$m(\widehat{\mathbf{w}}^*) = \min_{\mathbf{p} \in \Delta^n} \{\mathbf{p}^\top A \widehat{\mathbf{w}}^*\} - \frac{1}{2} \|\widehat{\mathbf{w}}^*\|_2^2 = \gamma^2 - \frac{\gamma^2}{2} = \frac{\gamma^2}{2}.$$

Thus, combining (20), (21), and Theorem 2, we have

$$m(\widehat{\mathbf{w}}^*) - m(\overline{\mathbf{w}}_T) \leq \overline{R}^{\mathbf{p}} + \overline{R}^{\mathbf{w}} \leq \frac{8 \log n}{T(T+1)} \leq \frac{8 \log n}{T^2}, \tag{22}$$

and thus

$$\begin{aligned}
\min_{\mathbf{p} \in \Delta^n} \mathbf{p}^\top A \overline{\mathbf{w}}_T &\geq m(\overline{\mathbf{w}}_T) \\
&\geq m(\widehat{\mathbf{w}}^*) - \frac{8 \log n}{T^2} \\
&= \frac{\gamma^2}{2} - \frac{8 \log n}{T^2},
\end{aligned} \tag{23}$$

where the first inequality follows because $-\frac{1}{2} \|\overline{\mathbf{w}}_T\|_2^2 \leq 0$. Hence, it can be seen that $\overline{\mathbf{w}}_T$ will have a non-negative margin when $T \geq \frac{4\sqrt{\log n}}{\gamma}$.

C.3 Proof of Proposition 2

We start from the original expression in Ji et al. (2021). For $\mathbf{g}_t$, we have

$$\begin{aligned}
\mathbf{g}_t &= \beta_t \left((-A^\top \mathbf{q}_t) + \mathbf{g}_{t-1} \right) \\
&= \frac{t}{t+1} \left((-A)^\top \mathbf{q}_t + \frac{t-1}{t} \mathbf{g}_{t-2} + \frac{t-1}{t} (-A)^\top \mathbf{q}_{t-1} \right) \\
&= \cdots \\
&= \frac{1}{t+1} (-A)^\top \left(\sum_{s=1}^{t} s \mathbf{q}_s \right).
\end{aligned} \tag{24}$$

Thus, we also have

$$\mathbf{g}_{t-1} = \frac{1}{t}(-A)^\top \left(\sum_{s=1}^{t-1} s\mathbf{q}_s \right).$$

Next, we focus on $\mathbf{v}_t$. Note that $\mathbf{v}_t = \mathbf{v}_{t-1} - \theta_{t-1}(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1})$, where

$$\begin{aligned} -\left(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1}\right) = \frac{1}{t} A^\top \left(\sum_{s=1}^{t-1} s\mathbf{q}_s + t\mathbf{q}_{t-1} \right) &= \frac{2(t+1)}{2t(t+1)} A^\top \left(\sum_{s=1}^{t-1} s\mathbf{q}_s + t\mathbf{q}_{t-1} \right) \\ &= \frac{t+1}{2}\mathbf{r}_t, \end{aligned} \tag{25}$$

where we define

$$\begin{aligned} \mathbf{r}_t = \frac{2}{t(t+1)} A^\top \left(\sum_{s=1}^{t-1} s\mathbf{q}_s + t\mathbf{q}_{t-1} \right) &= \frac{1}{\sum_{s=1}^t \alpha_s} A^\top \left(\sum_{s=1}^{t-1} s\mathbf{q}_s + t\mathbf{q}_{t-1} \right) \\ &= \operatorname*{argmin}_{\mathbf{r}\in\mathbb{R}^d} - \left(\sum_{s=1}^{t-1} s\mathbf{q}_s + t\mathbf{q}_{t-1} \right)^\top A\mathbf{r} + \frac{\sum_{s=1}^t \alpha_s}{2}\|\mathbf{r}\|_2^2 \\ &= \operatorname*{argmin}_{\mathbf{r}\in\mathbb{R}^d} \sum_{s=1}^{t-1} \alpha_s \left(-\mathbf{q}_s^\top A\mathbf{r} + \frac{1}{2}\|\mathbf{r}\|_2^2 \right) + \alpha_t \left(-\mathbf{q}_{t-1}^\top A\mathbf{r} + \frac{1}{2}\|\mathbf{r}\|_2^2 \right) \end{aligned} \tag{26}$$

As for the $\mathbf{q}_t$ variables, observe that

$$\begin{aligned} q_{t,i} &= \frac{\exp(-y^{(i)}\mathbf{v}_t^\top \mathbf{x}^{(i)})}{\sum_{j=1}^n \exp(-y^{(j)}\mathbf{v}_t^\top \mathbf{x}^{(j)})} \\ &= \frac{\exp(-y^{(i)}\mathbf{v}_{t-1}^\top \mathbf{x}^{(j)}) \exp(\frac{t}{2(t+1)}(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1})^\top y^{(i)}\mathbf{x}^{(i)})}{\sum_{j=1}^n \exp(-y^{(j)}\mathbf{v}_{t-1}^\top \mathbf{x}^{(j)}) \exp(\frac{t}{2(t+1)}(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1})^\top y^{(j)}\mathbf{x}^{(j)})} \\ &= \frac{\frac{\exp(-y^{(i)}\mathbf{v}_{t-1}^\top \mathbf{x}^{(j)})}{\sum_{k=1}^n \exp(-y^{(k)}\mathbf{v}_{t-1}^\top \mathbf{x}^{(k)})} \exp(\frac{t}{2(t+1)}(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1})^\top y^{(i)}\mathbf{x}^{(i)})}{\sum_{j=1}^n \frac{\exp(-y^{(j)}\mathbf{v}_{t-1}^\top \mathbf{x}^{(j)})}{\sum_{k=1}^n \exp(-y^{(k)}\mathbf{v}_{t-1}^\top \mathbf{x}^{(k)})} \exp(\frac{t}{2(t+1)}(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1})^\top y^{(j)}\mathbf{x}^{(j)})} \\ &= \frac{q_{t-1,i} \exp(\frac{t}{2(t+1)}(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1})^\top y^{(i)}\mathbf{x}^{(i)})}{\sum_{j=1}^n q_{t-1,j} \exp(\frac{t}{2(t+1)}(\mathbf{g}_{t-1} - A^\top \mathbf{q}_{t-1})^\top y^{(j)}\mathbf{x}^{(j)})} \\ &= \frac{q_{t-1,i} \exp(-\frac{\alpha_t}{4}\mathbf{r}_t^\top y^{(i)}\mathbf{x}^{(i)})}{\sum_{j=1}^n q_{t-1,j} \exp(-\frac{\alpha_t}{4}\mathbf{r}_t^\top y^{(j)}\mathbf{x}^{(j)})} = \frac{q_{0,i} \exp(-\sum_{s=1}^t \frac{\alpha_s}{4}\mathbf{r}_s^\top y^{(i)}\mathbf{x}^{(i)})}{\sum_{j=1}^n q_{0,j} \exp(-\sum_{s=1}^t \frac{\alpha_s}{4}\mathbf{r}_s^\top y^{(j)}\mathbf{x}^{(j)})}. \end{aligned} \tag{27}$$

Based on the relationship between the multiplicative weights algorithm and FTRL with the entropy regularizer, we have

$$\mathbf{q}_t = \operatorname*{argmin}_{\mathbf{q}\in\Delta^n} \sum_{s=1}^t \frac{\alpha_s}{4}\mathbf{q}^\top A\mathbf{r}_s + \mathcal{D}_E(\mathbf{q}, \mathbf{q}_0). \tag{28}$$

Next, note that $\mathbf{r}_1 = A^\top \mathbf{q}_0 = \frac{A^\top \mathbf{1}}{n} = \mathbf{w}_1$, from the proof of Proposition 1. Combining this with the fact that $\mathbf{q}_0 = \frac{\mathbf{1}}{n} = \mathbf{p}_0$, we can recursively conclude that $\mathbf{r}_t = \mathbf{w}_t$ and $\mathbf{q}_t = \mathbf{p}_t$. Finally, since $\mathbf{v}_0 = \mathbf{0}$, we have that

$$\mathbf{v}_T = \mathbf{v}_{T-1} + \theta_{T-1}\frac{T+1}{2}\mathbf{r}_T = \mathbf{v}_{T-1} + \frac{T}{2(T+1)}\frac{T+1}{2}\mathbf{r}_T = \mathbf{v}_{T-1} + \frac{1}{4}\alpha_T\mathbf{r}_T = \frac{1}{4}\sum_{t=1}^T \alpha_t\mathbf{r}_t$$

C.4 Proof of Theorem 4

For the first two parts of Theorem 4, note that the expressions in Algorithms 1 and and 2 are exactly the same. In the following, we focus on the normalized margin analysis. First, recall that the

normalized margin is defined as:

$$\overline{\gamma}(\mathbf{w}) = \frac{\min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{w}}{\|\mathbf{w}\|_2},$$

and for the the maximal margin classifier $\mathbf{w}^*$ we have $\overline{\gamma}(\mathbf{w}^*) = \gamma$. Next, we compute $\overline{\gamma}(\mathbf{v}_T)$ under our framework. Note that

$$\overline{\mathbf{w}}_T = \frac{4\mathbf{v}_T}{\sum_{t=1}^T \alpha_t} = \frac{8\mathbf{v}_T}{T(T+1)}.$$

Moreover, from the proof of Theorem 3, we know that $m(\mathbf{w}) - m(\overline{\mathbf{w}}_T) \leq \frac{8\log n}{T(T+1)}$ for any $\mathbf{w} \in \mathbb{R}^d$. Thus, we have

$$\begin{aligned}
\min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\frac{8\mathbf{v}_T}{T(T+1)} - \frac{1}{2}\left\|\frac{8\mathbf{v}_T}{T(T+1)}\right\|_2^2 &= \min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\overline{\mathbf{w}}_T - \frac{1}{2}\|\overline{\mathbf{w}}_T\|_2^2 \\
&= m(\overline{\mathbf{w}}_T) \\
&\geq m\left(\frac{8}{T(T+1)}\|\mathbf{v}_T\|_2\mathbf{w}^*\right) - \frac{8\log n}{T(T+1)} \\
&= \frac{8}{T(T+1)}\|\mathbf{v}_T\|_2\gamma - \frac{64}{2(T(T+1))^2}\|\mathbf{v}_T\|_2^2 - \frac{8\log n}{T(T+1)}
\end{aligned}$$

Rearranging and multiplying both sides by $\frac{T(T+1)}{8\|\mathbf{v}_T\|_2}$, we obtain

$$\frac{\min_{\mathbf{p}} \mathbf{p}^\top A\mathbf{v}_T}{\|\mathbf{v}_T\|_2} \geq \gamma - \frac{\log n}{\|\mathbf{v}_T\|_2}$$

The proof is finished by the following lower bound on $\|\mathbf{v}_T\|_2$.

Lemma 5. *We have*

$$\|\mathbf{v}_T\|_2 \geq \frac{T(T+1)\gamma}{8}.$$

Proof. Note that $\|\mathbf{w}^*\| = 1$, so $\frac{\mathbf{v}_T^\top \mathbf{w}^*}{\|\mathbf{v}_T\|_2} \leq 1$. Then, based on our derivation of $\mathbf{w}_t$ in the proof of Theorem 3, we can conclude that

$$\begin{aligned}
\|\mathbf{v}_T\|_2 \geq \mathbf{v}_T^\top \mathbf{w}^* &= \frac{1}{4}\sum_{t=1}^T \alpha_t \mathbf{w}_t^\top \mathbf{w}^* \\
&= \frac{1}{4}\sum_{t=1}^T \alpha_t \left[\frac{1}{\sum_{s=1}^t \alpha_s}\left(\sum_{s=1}^{t-1}\alpha_s\mathbf{p}_s + \alpha_t\mathbf{p}_t\right)\right]^\top A\mathbf{w}^* \\
&\geq \frac{1}{4}\sum_{t=1}^T \alpha_t \min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{w}^* \\
&= \frac{T(T+1)\gamma}{8}.
\end{aligned} \tag{29}$$

□

C.5 Details of Section 4.4

The Perceptron proposed by Yu et al. (2014), named mirror-prox algorithm for feasibility Problems (MPFP) procedure (Yu et al., 2014), is summarized in Algorithm 5. Yu et al. (2014) consider a general objective $g : \mathcal{W} \times \mathcal{P} \to \mathbb{R}$, for spaces $\mathcal{P} \subset \mathbb{R}^n$ and $\mathcal{W} \subset \mathbb{R}^d$, and defines $F(\mathbf{w}, \mathbf{p}) = [-\nabla_{\mathbf{w}} g(\mathbf{w}, \mathbf{p}); \nabla_{\mathbf{p}} g(\mathbf{w}, \mathbf{p})] \in \mathbb{R}^{d+n}$ where $[\mathbf{x}; \mathbf{y}]$ denotes the concatenation of vectors $\mathbf{x}$ and $\mathbf{y}$. Associated to $\mathcal{P}$ (and analogously for $\mathcal{W}$), Yu et al. (2014) make use of the following quantities:

Algorithm 5 MPFP of Yu et al. (2014)

Input: $\{\gamma_t\}_{t=1}^T, \mathbf{v}_1 = \mathbf{r}^*$
for $t = 1, \dots, T$ **do**
  $\mathbf{u}_t = \text{Prox}_{\mathbf{v}_t}^R(\gamma_t F(\mathbf{v}_t))$
  $\mathbf{v}_{t+1} = \text{Prox}_{\mathbf{v}_t}^R(\gamma_t F(\mathbf{u}_t))$
  $\mathbf{z}_t = \frac{1}{\sum_{s=1}^t \gamma_s} \sum_{s=1}^t \gamma_s \mathbf{u}_s$
end for
Output: $\mathbf{z}_T$

$$
\begin{aligned}
\psi_t(\mathbf{w}) &= -\widehat{\mathbf{p}}_{t-1}^\top A\mathbf{w} \\
\phi_t(\mathbf{p}) &= \mathbf{p}^\top A\widehat{\mathbf{w}}_{t-1} \\
\mathsf{OL}^{\mathbf{w}} = \text{OMD}\left[\frac{1}{2}\|\cdot\|_2^2, \mathbf{0}, \frac{1}{\sqrt{\log n}}, \psi_t, \mathcal{B}_d\right] \Leftrightarrow \mathbf{w}_t &= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{B}_d} \frac{1}{\sqrt{\log n}} \langle \nabla\psi_t(\mathbf{w}_{t-1}), \mathbf{w}\rangle \\
&\quad + \mathcal{D}_{\frac{1}{2}\|\cdot\|_2^2}(\mathbf{w}, \widehat{\mathbf{w}}_{t-1}) \\
\widehat{\mathbf{w}}_t &= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{B}_d} \frac{1}{\sqrt{\log n}} \langle \nabla h_t(\mathbf{w}_t), \mathbf{w}\rangle \\
&\quad + \mathcal{D}_{\frac{1}{2}\|\cdot\|_2^2}(\mathbf{w}, \widehat{\mathbf{w}}_{t-1}) \\
\mathsf{OL}^{\mathbf{p}} = \text{OMD}\left[E, \frac{\mathbf{1}}{n}, \sqrt{\log n}, \phi_t, \Delta^n\right] \Leftrightarrow \mathbf{p}_t &= \operatorname*{argmin}_{\mathbf{p}\in\Delta^n} \sqrt{\log n}\, \langle \nabla\phi_t(\mathbf{p}_{t-1}), \mathbf{p}\rangle \\
&\quad + \mathcal{D}_E(\mathbf{p}, \widehat{\mathbf{p}}_{t-1}) \\
\widehat{\mathbf{p}}_t &= \operatorname*{argmin}_{\mathbf{p}\in\Delta^n} \sqrt{\log n}\, \langle \nabla\ell_t(\mathbf{p}_{t-1}), \mathbf{p}\rangle \\
&\quad + \mathcal{D}_E(\mathbf{p}, \widehat{\mathbf{p}}_{t-1})
\end{aligned}
$$

- A constant $\alpha_{\mathbf{p}} > 0$ and function $R_{\mathbf{p}} : \mathcal{P} \to \mathbb{R}$ that is 1-strongly convex w.r.t. some norm $\|\cdot\|_{\mathcal{P}}$ on $\mathcal{P}$;
- Minimizer $\mathbf{r}_{\mathbf{p}}^* = \text{argmin}_{\mathbf{p}\in\mathcal{P}} R_{\mathbf{p}}(\mathbf{p})$.
- Bregman divergence $D_{\mathbf{p}}(\mathbf{p}, \mathbf{q}) = D_{R_{\mathbf{p}}}(\mathbf{p}, \mathbf{q})$;
- Diameter $\Omega_{\mathbf{p}} = \max_{\mathbf{p}\in\mathcal{P}} D_{\mathbf{p}}(\mathbf{p}, \mathbf{r}_{\mathbf{p}}^*)$.

From these, we can define the function $R : \mathbb{R}^{d+n} \mapsto \mathbb{R}$ and its minimizer

$$R(\mathbf{w}, \mathbf{p}) = \alpha_{\mathbf{w}} R_{\mathbf{w}}(\mathbf{w}) + \alpha_{\mathbf{p}} R_{\mathbf{p}}(\mathbf{p}) \quad \text{and} \quad \mathbf{r}^* = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^{d+n}} R(\mathbf{x}) = [\mathbf{r}_{\mathbf{w}}^*; \mathbf{r}_{\mathbf{p}}^*]$$

Lastly, the algorithm relies on the following Prox map: for function $f : \mathcal{Z} \to \mathbb{R}$ and point $\mathbf{v} \in \mathcal{Z}$ on a general space $\mathcal{Z} \subset \mathbb{R}^m$, define the mapping $\text{Prox}_{\mathbf{v}}^f : \mathcal{Z} \to \mathcal{Z}$ by

$$\text{Prox}_{\mathbf{v}}^f(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{u}\in\mathcal{Z}} \left\{\mathbf{u}^\top \mathbf{z} + D_f(\mathbf{u}, \mathbf{v})\right\}$$

In this work, we consider MPFP under the bilinear objective $g(\mathbf{w}, \mathbf{p}) = \mathbf{p}^\top A\mathbf{w}$, where the rows of A are assumed to be normalized w.r.t. $\|\cdot\|_2$. We further give the following specifications:

- Spaces $\mathcal{P} = \Delta_n$ and $\mathcal{W} = \mathcal{B}_d = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\|_2 \le 1\}$.
- Norms $\|\cdot\|_{\mathcal{P}} = \|\cdot\|_1$ and $\|\cdot\|_{\mathcal{W}} = \|\cdot\|_2$.
- Regularizers $R_{\mathbf{p}} = E$, where E denotes the entropy regularizer (i.e., negative Shannon entropy), and $R_{\mathbf{w}} = \frac{1}{2}\|\cdot\|_2^2$.
- Contants $\alpha_{\mathbf{p}} = \frac{1}{\sqrt{\Omega_{\mathbf{p}}}(\sqrt{\Omega_{\mathbf{p}}}+\sqrt{\Omega_{\mathbf{w}}})}$ and $\alpha_{\mathbf{w}} = \frac{\sqrt{\Omega_{\mathbf{p}}}}{\sqrt{\Omega_{\mathbf{p}}}+\sqrt{\Omega_{\mathbf{w}}}}$.

- Constant weights $\gamma_t = \frac{1}{\sqrt{\Omega_{\mathbf{p}}}+\sqrt{\Omega_{\mathbf{w}}}}$ for each $t \in [T]$.

Endowed with these parameters, the procedure becomes the MPLFP algorithm of Yu et al. (2014), and we can conclude the following connection.

Theorem 8. *Let $\alpha_t = 1$. Then the MPLFP algorithm is equivalent to the game dynamics given by Algorithm 5, in the sense that $\mathbf{u}_t = [\mathbf{w}_t; \mathbf{p}_t]$, $\mathbf{v}_{t+1} = [\widehat{\mathbf{w}}_t; \widehat{\mathbf{p}}_t]$, and $\mathbf{z}_t = \left[\frac{1}{t}\sum_{s=1}^{t} \mathbf{w}_s; \frac{1}{t}\sum_{s=1}^{t} \mathbf{p}_s\right]$.*

Proof. We begin by noting that

$$\Omega_{\mathbf{w}} = \max_{\mathbf{w}\in\mathcal{B}_d} D_{\mathbf{w}}(\mathbf{w}, \mathbf{r}_w^*) = \max_{\mathbf{w}\in\mathcal{B}_d} \frac{1}{2}\|\mathbf{w}\|_2^2 = \frac{1}{2}$$

$$\Omega\mathbf{p} = \max_{\mathbf{p}\in\Delta^n} D_{\mathbf{p}}(\mathbf{p}, \mathbf{r}_p^*) = \max_{\mathbf{p}\in\Delta^n} \mathrm{KL}\left(\mathbf{p}||\frac{\mathbf{1}}{n}\right) = \log n$$

where KL denotes the Kullback–Leibler divergence, so that we can derive parameters

$$\gamma_t = \frac{1}{\sqrt{\log n} + \frac{1}{\sqrt{2}}} \quad \text{and} \quad \alpha_{\mathbf{p}} = \frac{1}{\sqrt{\log n}\left(\sqrt{\log n} + \frac{1}{\sqrt{2}}\right)} \quad \text{and} \quad \alpha_{\mathbf{w}} = \frac{\sqrt{\log n}}{\sqrt{\log n} + \frac{1}{\sqrt{2}}}$$

Next, note that $F(\mathbf{w}, \mathbf{p}) = \left[-\mathbf{p}^\top A; A\mathbf{w}\right]$. For convenience, we will write the MPFP iterates as

$$\mathbf{u}_t = [\mathbf{x}_t; \mathbf{y}_t] \quad \text{and} \quad \mathbf{v}_{t+1} = [\widehat{\mathbf{x}}_t; \widehat{\mathbf{y}}_t]$$

Section 10 of Yu et al. (2014) shows that

$$\mathrm{Prox}^R_{[\widehat{\mathbf{x}};\widehat{\mathbf{y}}]}(\mathbf{x}, \mathbf{y}) = \left[\mathrm{Prox}^{R_{\mathbf{w}}}_{\widehat{\mathbf{x}}}\left(\frac{\mathbf{x}}{\alpha_{\mathbf{w}}}\right); \mathrm{Prox}^{R_{\mathbf{p}}}_{\widehat{\mathbf{y}}}\left(\frac{\mathbf{y}}{\alpha_{\mathbf{p}}}\right)\right]$$

Using the iterate definitions, we can then obtain

$$\begin{aligned}
[\mathbf{x}_t; \mathbf{y}_t] &= \mathbf{u}_t \\
&= \mathrm{Prox}^R_{\mathbf{v}_t}(\gamma_t F(\mathbf{v}_t)) \\
&= \mathrm{Prox}^R_{[\widehat{\mathbf{x}}_{t-1};\widehat{\mathbf{y}}_{t-1}]}(-\gamma_t \widehat{\mathbf{y}}_{t-1}^\top A, \gamma_t A\widehat{\mathbf{x}}_{t-1}) \\
&= \left[\mathrm{Prox}^{R_{\mathbf{w}}}_{\widehat{\mathbf{x}}_{t-1}}\left(\frac{-\gamma_t \widehat{\mathbf{y}}_{t-1}^\top A}{\alpha_{\mathbf{w}}}\right); \mathrm{Prox}^{R_{\mathbf{p}}}_{\widehat{\mathbf{y}}_{t-1}}\left(\frac{\gamma_t A\widehat{\mathbf{x}}_{t-1}}{\alpha_{\mathbf{p}}}\right)\right] \\
&= \left[\mathrm{Prox}^{R_{\mathbf{w}}}_{\widehat{\mathbf{x}}_{t-1}}\left(-\frac{1}{\sqrt{\log n}}\widehat{\mathbf{y}}_{t-1}^\top A\right); \mathrm{Prox}^{R_{\mathbf{p}}}_{\widehat{\mathbf{y}}_{t-1}}\left(\sqrt{\log n} A\widehat{\mathbf{x}}_{t-1}\right)\right]
\end{aligned}$$

and

$$\begin{aligned}
[\widehat{\mathbf{x}}_t; \widehat{\mathbf{y}}_t] &= \mathbf{v}_{t+1} \\
&= \mathrm{Prox}^R_{\mathbf{v}_t}(\gamma_t F(\mathbf{u}_t)) \\
&= \mathrm{Prox}^R_{[\widehat{\mathbf{x}}_{t-1};\widehat{\mathbf{y}}_{t-1}]}(-\gamma_t \mathbf{y}_t^\top A, \gamma_t A\mathbf{x}_t) \\
&= \left[\mathrm{Prox}^{R_{\mathbf{w}}}_{\widehat{\mathbf{x}}_{t-1}}\left(\frac{-\gamma_t \mathbf{y}_t^\top A}{\alpha_{\mathbf{w}}}\right); \mathrm{Prox}^{R_{\mathbf{p}}}_{\widehat{\mathbf{y}}_{t-1}}\left(\frac{\gamma_t A\mathbf{x}_t}{\alpha_{\mathbf{p}}}\right)\right] \\
&= \left[\mathrm{Prox}^{R_{\mathbf{w}}}_{\widehat{\mathbf{x}}_{t-1}}\left(-\frac{1}{\sqrt{\log n}}\mathbf{y}_t^\top A\right); \mathrm{Prox}^{R_{\mathbf{p}}}_{\widehat{\mathbf{y}}_{t-1}}\left(\sqrt{\log n} A\mathbf{x}_t\right)\right]
\end{aligned}$$

Using the definition of the Prox function then yields

$$\begin{aligned}
\mathbf{x}_t &= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{B}_d}\left\{-\frac{1}{\sqrt{\log n}}\widehat{\mathbf{y}}_{t-1}^\top A\mathbf{w} + D_{\mathbf{w}}(\mathbf{w}, \widehat{\mathbf{x}}_{t-1})\right\} \\
\widehat{\mathbf{x}}_t &= \operatorname*{argmin}_{\mathbf{w}\in\mathcal{B}_d}\left\{-\frac{1}{\sqrt{\log n}}\mathbf{y}_t^\top A\mathbf{w} + D_{\mathbf{w}}(\mathbf{w}, \widehat{\mathbf{x}}_{t-1})\right\} \\
\mathbf{y}_t &= \operatorname*{argmin}_{\mathbf{p}\in\Delta^n}\left\{\sqrt{\log n}\mathbf{p}^\top A\widehat{\mathbf{x}}_{t-1} + D_{\mathbf{p}}(\mathbf{p}, \widehat{\mathbf{y}}_{t-1})\right\} \\
\widehat{\mathbf{y}}_t &= \operatorname*{argmin}_{\mathbf{p}\in\Delta^n}\left\{\sqrt{\log n}\mathbf{p}^\top A\mathbf{x}_t + D_{\mathbf{p}}(\mathbf{p}, \widehat{\mathbf{y}}_{t-1})\right\}
\end{aligned}$$

From the fact that

$$[\widehat{\mathbf{x}}_0; \widehat{\mathbf{y}}_0] = v_1 = [\mathbf{r}^*_{\mathbf{w}}; \mathbf{r}^*_{\mathbf{p}}] = \left[\mathbf{0}; \frac{\mathbf{1}}{n}\right] = [\widehat{\mathbf{w}}_0; \widehat{\mathbf{p}}_0]$$

and plugging in the appropriate parameters into the OMD updates, we can recursively conclude that $[\widehat{\mathbf{w}}_t; \widehat{\mathbf{p}}_t] = [\widehat{\mathbf{x}}_t; \widehat{\mathbf{y}}_t] = \mathbf{v}_{t+1}$ and $[\mathbf{w}_t; \mathbf{p}_t] = [\mathbf{x}_t; \mathbf{y}_t] = \mathbf{u}_t$ for each $t \in [T]$. As a result, we also have that

$$\begin{aligned}
\mathbf{z}_t &= \frac{1}{\sum_{s=1}^t \gamma_s} \sum_{s=1}^t \gamma_s \mathbf{u}_s \\
&= \left[\frac{1}{\sum_{s=1}^t \gamma_s} \sum_{s=1}^t \gamma_s \mathbf{x}_s; \frac{1}{\sum_{s=1}^t \gamma_s} \sum_{s=1}^t \gamma_s \mathbf{y}_s\right] \\
&= \left[\frac{1}{t} \sum_{s=1}^t \mathbf{w}_s; \frac{1}{t} \sum_{s=1}^t \mathbf{p}_s\right]
\end{aligned}$$

□

Proposition 3. *Let $\alpha_t = 1$. Then the total regret of both players under the dynamics of Algorithm 5 is bounded by*

$$R^{\mathbf{p}} + R^{\mathbf{w}} = O(\sqrt{\log n})$$

Before we prove this result, we need to state a technical tool that we will make use of.

Lemma 6. *For any $\mathbf{p} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^d$, we have that*

$$\|\mathbf{p}^\top A\|_2 \leq \|\mathbf{p}\|_1 \quad \textit{and} \quad \|A\mathbf{w}\|_\infty \leq \|\mathbf{w}\|_2$$

Proof. For the first inequality, let $\tilde{\mathbf{p}} = \frac{\mathbf{p}}{\|\mathbf{p}\|_1}$. Then,

$$\|\mathbf{p}^\top A\|_2 = \|\mathbf{p}\|_1 \|\tilde{\mathbf{p}}^\top A\|_2 = \|\mathbf{p}\|_1 \left\|\sum_{i=1}^n \tilde{p}_i A_{(i,:)}\right\|_2 \leq \|\mathbf{p}\|_1 \sum_{i=1}^n |\tilde{p}_i| \|A_{(i,:)}\|_2 = \|\mathbf{p}\|_1$$

For the second inequality, we have that

$$\|A\mathbf{w}\|_\infty = \max_{i \in [n]} \left|A^\top_{(i,:)} \mathbf{w}\right| \leq \|\mathbf{w}\|_2 \max_{i \in [n]} \|A_{(i,:)}\|_2 = \|\mathbf{w}\|_2$$

□

Proof of Proposition 3. Let us first note that, from the proof of Theorem 8, we know that

$$\Omega_{\mathbf{w}} = \max_{\mathbf{w} \in \mathcal{B}_d} D_{\mathbf{w}}(\mathbf{w}, \widehat{\mathbf{w}}_0) = \frac{1}{2} \quad \text{and} \quad \Omega_{\mathbf{p}} = \max_{\mathbf{p} \in \Delta^n} D_{\mathbf{p}}(\mathbf{p}, \widehat{\mathbf{p}}_0) = \log n$$

Then, using Lemmas 4 and 6, we can bound the regret of the $\mathbf{w}$-player as follows:

$$\begin{aligned}
R^{\mathbf{w}} &\leq \frac{\sqrt{\log n}}{2} + \sum_{t=1}^T \left[\|(\mathbf{p}_t - \widehat{\mathbf{p}}_{t-1})^\top A\|_2 \|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|_2 - \frac{\sqrt{\log n}}{2} \left(\|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|_2^2 + \|\mathbf{w}_t - \widehat{\mathbf{w}}_{t-1}\|_2^2\right)\right] \\
&\leq \frac{\sqrt{\log n}}{2} + \sum_{t=1}^T \left[\|\mathbf{p}_t - \widehat{\mathbf{p}}_{t-1}\|_1 \|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|_2 - \frac{\sqrt{\log n}}{2} \left(\|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|_2^2 + \|\mathbf{w}_t - \widehat{\mathbf{w}}_{t-1}\|_2^2\right)\right]
\end{aligned}$$

Similarly, for the $\mathbf{p}$-player we obtain

$$\begin{aligned}
R^{\mathbf{p}} &\leq \sqrt{\log n} + \sum_{t=1}^T \left[\|A(\mathbf{w}_t - \widehat{\mathbf{w}}_{t-1})\|_\infty \|\mathbf{p}_t - \widehat{\mathbf{p}}_t\|_1 - \frac{1}{2\sqrt{\log n}} \left(\|\mathbf{p}_t - \widehat{\mathbf{p}}_t\|_1^2 + \|\mathbf{p}_t - \widehat{\mathbf{p}}_{t-1}\|_1^2\right)\right] \\
&\leq \sqrt{\log n} + \sum_{t=1}^T \left[\|\mathbf{w}_t - \widehat{\mathbf{w}}_{t-1}\|_2 \|\mathbf{p}_t - \widehat{\mathbf{p}}_t\|_1 - \frac{1}{2\sqrt{\log n}} \left(\|\mathbf{p}_t - \widehat{\mathbf{p}}_t\|_1^2 + \|\mathbf{p}_t - \widehat{\mathbf{p}}_{t-1}\|_1^2\right)\right]
\end{aligned}$$

Adding both bounds then yields

$$\begin{aligned}
R^{\mathbf{w}} + R^{\mathbf{p}} &\leq \frac{\sqrt{\log n}}{2} + \sqrt{\log n} \\
&\quad - \frac{1}{2}\sum_{t=1}^{T}\left[\frac{1}{\sqrt{\log n}}\|\mathbf{p}_t - \widehat{\mathbf{p}}_{t-1}\|_1^2 + \sqrt{\log n}\|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|_2^2 - 2\|\mathbf{p}_t - \widehat{\mathbf{p}}_{t-1}\|_1\|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|_2\right. \\
&\qquad \left. + \sqrt{\log n}\|\mathbf{w}_t - \widehat{\mathbf{w}}_{t-1}\|_2^2 + \frac{1}{\sqrt{\log n}}\|\mathbf{p}_t - \widehat{\mathbf{p}}_t\|_1^2 - 2\|\mathbf{w}_t - \widehat{\mathbf{w}}_{t-1}\|_2\|\mathbf{p}_t - \widehat{\mathbf{p}}_t\|_1\right] \\
&= \frac{\sqrt{\log n}}{2} + \sqrt{\log n} \\
&\quad - \frac{1}{2}\sum_{t=1}^{T}\left[\left((\log n)^{-\frac{1}{4}}\|\mathbf{p}_t - \widehat{\mathbf{p}}_{t-1}\|_1 - (\log n)^{\frac{1}{4}}\|\mathbf{w}_t - \widehat{\mathbf{w}}_t\|_2\right)^2\right. \\
&\qquad \left.\left((\log n)^{\frac{1}{4}}\|\mathbf{w}_t - \widehat{\mathbf{w}}_{t-1}\|_2 - (\log n)^{-\frac{1}{4}}\|\mathbf{p}_t - \widehat{\mathbf{p}}_t\|_1\right)^2\right] \\
&\leq \frac{\sqrt{\log n}}{2} + \sqrt{\log n}
\end{aligned}$$

□

A nice property of the MPLFP algorithm is that it provides a linear separator when the data is linearly separable and an approximate certificate of infeasibility when it is not. Such a certificate is a vector $\widehat{\mathbf{p}} \in \Delta^n$ satisfying $\widehat{\mathbf{p}}^\top A = \mathbf{0}$; the duality of the two problems can be observed via Gordan's Theorem. We call $\widehat{\mathbf{p}} \in \Delta^n$ an ϵ-certificate of infeasibility if $\|\widehat{\mathbf{p}}A\|_2 \leq \epsilon$. Next, we recover the convergence analysis of MPLFP, from Yu et al. (2014), using the tools of no-regret learning.

Theorem 9. *Let $\alpha_t = 1$. Then under Assumption 1, $\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$ has a non-negative margin when $T = \Omega\left(\frac{\sqrt{\log n}}{\gamma}\right)$. On the other hand, if $\max_{\mathbf{w}\in\mathcal{B}_d}\gamma(\mathbf{w}) < 0$, then $\frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_t$ is an ϵ-certificate of infeasibility when $T = \Omega\left(\frac{\sqrt{\log n}}{\epsilon}\right)$.*

Proof. Note that in this setting, we have that $m(\mathbf{w}) = \gamma(\mathbf{w})$ for each $\mathbf{w} \in \mathcal{B}_d$, and $\frac{1}{T}\sum_{t=1}^{T} w_t \in \mathcal{B}_d$. Then, under Assumption 1, we have that $\mathbf{w}^* \in \mathcal{B}_d$, so Theorem 2 and Proposition 3 yield

$$\gamma\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) = m\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) \geq m(\mathbf{w}^*) - \frac{R^{\mathbf{w}} + R^{\mathbf{p}}}{T} = \gamma(\mathbf{w}^*) - \frac{R^{\mathbf{w}} + R^{\mathbf{p}}}{T} \geq \gamma - \frac{3\sqrt{\log n}}{2T}$$

That is, the margin is non-negative provided that $T \geq \frac{3\sqrt{\log n}}{2\gamma}$.

To show the second conclusion, let us begin by noting that with a slight modification to the proof of Theorem 2, we can obtain

$$\max_{\mathbf{w}\in\mathcal{B}_d} g\left(\mathbf{w}, \frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_t\right) \leq \min_{\mathbf{p}\in\Delta^n} g\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t, \mathbf{p}\right) + \bar{R}^{\mathbf{w}} + \bar{R}^{\mathbf{p}} = m\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) + \bar{R}^{\mathbf{w}} + \bar{R}^{\mathbf{p}}$$

Our assumption now is that $\max_{\mathbf{w}\in\mathcal{B}_d} m(\mathbf{w}) = \max_{\mathbf{w}\in\mathcal{B}_d}\gamma(\mathbf{w}) < 0$, so that

$$\begin{aligned}
\left\|\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_t\right)^\top A\right\|_2 &= \max_{\mathbf{w}\in\mathcal{B}_d}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_t\right)^\top A\mathbf{w} && \text{Cauchy-Schwarz} \\
&= \max_{\mathbf{w}\in\mathcal{B}_d} g\left(\mathbf{w}, \frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_t\right) \\
&< \bar{R}^{\mathbf{w}} + \bar{R}^{\mathbf{p}} \\
&\leq \frac{3\sqrt{\log n}}{2T}
\end{aligned}$$

Hence, $\frac{1}{T}\sum_{t=1}^{T}\mathbf{p}_t$ is an ϵ-certificate of infeasibility provided that $T \geq \frac{3\sqrt{\log n}}{2\epsilon}$. □

D PROOF OF SECTION 5

D.1 PROOF OF THEOREM 6

First, based on the property of the objective function from Ji et al. (2021), we have

$$-A^\top \mathbf{q}_t = \frac{\nabla R(\mathbf{u}_t)}{R(\mathbf{u}_t)},$$

where

$$q_{t,i} = \frac{\exp\left(-y^{(i)}\mathbf{u}_t^\top \mathbf{x}^{(i)}\right)}{\sum_{j=1}^n \exp\left(-y^{(j)}\mathbf{u}_t^\top \mathbf{x}^{(j)}\right)}. \tag{30}$$

Then, using the fact that $\mathbf{v}_0 = \mathbf{0}$, we obtain

$$\mathbf{v}_t = \mathbf{v}_{t-1} + tA^\top \mathbf{q}_t = \sum_{t=1}^T \alpha_t A^\top \mathbf{q}_t = \left(\sum_{i=1}^t \alpha_i\right) \cdot \mathbf{r}_t,$$

where we define $\mathbf{r}_0 = 0$ and $\mathbf{r}_t = \frac{1}{\sum_{i=1}^t \alpha_i} \sum_{i=1}^t \alpha_i A^\top \mathbf{q}_i$ for $t \geq 1$. Then, we can rewrite

$$\mathbf{s}_t = \mathbf{s}_{t-1} + \frac{1}{2(t+1)}\mathbf{v}_t = \sum_{i=1}^t \frac{1}{2(i+1)}\mathbf{v}_i = \sum_{i=1}^t \frac{1}{2i(i+1)} i\mathbf{v}_i = \frac{1}{4}\sum_{i=1}^t \frac{1}{\sum_{j=1}^i \alpha_j} i\mathbf{v}_i = \frac{1}{4}\sum_{i=1}^t \alpha_i \mathbf{r}_i.$$

where we used the initialization $\mathbf{s}_0 = \mathbf{0}$. Moreover,

$$\mathbf{u}_t = \mathbf{s}_{t-1} + \frac{1}{2(t-1)}\mathbf{v}_{t-1} = \frac{1}{4}\sum_{i=1}^{t-1} \alpha_i \mathbf{r}_i + \frac{1}{2(t-1)}\mathbf{v}_{t-1} = \frac{1}{4}\sum_{i=1}^{t-1} \alpha_i \mathbf{r}_i + \frac{1}{4}\alpha_t \mathbf{r}_{t-1}. \tag{31}$$

Combining (30) and (31), we have

$$\begin{aligned}
\mathbf{q}_t &= \operatorname*{argmin}_{\mathbf{q}\in\Delta^n} \mathbf{q}^\top A\mathbf{u}_t + D_E\left(\mathbf{q}, \frac{\mathbf{1}}{n}\right) \\
&= \operatorname*{argmin}_{\mathbf{q}\in\Delta^n} \frac{1}{4}\left(\sum_{i=1}^{t-1} \alpha_i \mathbf{q}^\top A\mathbf{r}_i + \alpha_t \mathbf{q}^\top A\mathbf{r}_{t-1}\right) + D_E\left(\mathbf{q}, \frac{\mathbf{1}}{n}\right).
\end{aligned}$$

Additionally,

$$\begin{aligned}
\mathbf{r}_t = \frac{1}{\sum_{i=1}^t \alpha_i}\sum_{i=1}^t \alpha_i A^\top \mathbf{q}_i &= \operatorname*{argmin}_{\mathbf{r}\in\mathbb{R}^d} \sum_{i=1}^t \alpha_i \mathbf{q}_i^\top A\mathbf{r} + \frac{\sum_{i=1}^t \alpha_i}{2}\|\mathbf{r}\|_2^2 \\
&= \operatorname*{argmin}_{\mathbf{r}\in\mathbb{R}^d} \sum_{i=1}^t \alpha_i \left(\mathbf{q}_i^\top A\mathbf{r} + \frac{1}{2}\|\mathbf{r}\|_2^2\right).
\end{aligned} \tag{32}$$

Note that since $\mathbf{r}_0 = \mathbf{w}_0$ and we can drop the $\frac{1}{2}\|\cdot\|_2^2$ terms from the minimization in the $\mathbf{p}$-player's update (since it doesn't depend on $\mathbf{p}$), we get that $\mathbf{q}_t = \mathbf{p}_t$ and $\mathbf{r}_t = \mathbf{w}_t$. Finally, we can conclude that

$$\mathbf{s}_T = \frac{1}{4}\sum_{t=1}^T \alpha_t \mathbf{r}_t = \frac{\sum_{t=1}^T \alpha_t}{4}\frac{1}{\sum_{t=1}^T \alpha_t}\sum_{t=1}^T \alpha_t \mathbf{w}_t = \frac{\sum_{t=1}^T \alpha_t}{4}\overline{\mathbf{w}}_T = \widetilde{\mathbf{w}}_T$$

Next, we turn to the regret. For the $\mathbf{p}$-player, it uses OFTRL, so, based on Lemma 3, we obtain

$$\begin{aligned}
&\sum_{t=1}^T \alpha_t \ell_t(\mathbf{p}_t) - \sum_{t=1}^T \alpha_t \ell_t(\mathbf{p}) \\
&\leq 4\left(E(\mathbf{p}) - E\left(\frac{\mathbf{1}}{n}\right)\right) + \sum_{t=1}^T \frac{t^2\|A\mathbf{w}_t - A\mathbf{w}_{t-1}\|_\infty^2}{8} \\
&\leq 4\log n + \sum_{t=1}^T \frac{t^2\left[\max_{i\in[n]}\left|y^{(i)}(\mathbf{w}_t - \mathbf{w}_{t-1})^\top \mathbf{x}^{(i)}\right|\right]\left[\max_{i\in[n]}\left|y^{(i)}(\mathbf{w}_t - \mathbf{w}_{t-1})^\top \mathbf{x}^{(i)}\right|\right]}{8} \\
&\leq 4\log n + \sum_{t=1}^T \frac{t^2\|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2}{8}.
\end{aligned} \tag{33}$$

On the other hand, the $\mathbf{w}$-player applies FTRL$^+$, and note that $h_t(\mathbf{w})$ is 1-strongly convex. Thus, the regret is bounded by

$$\sum_{t=1}^{T} \alpha_t h_t(\mathbf{w}_t) - \sum_{t=1}^{T} \alpha_t h_t(\mathbf{w}) \leq -\sum_{t=1}^{T} \frac{t(t-1)}{4} \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2. \tag{34}$$

Finally, we focus on the margin. Combining (33), and (34), we have

$$\overline{R}^{\mathbf{w}} + \overline{R}^{\mathbf{p}} = \frac{2}{T(T+1)} \left(4\log n + \sum_{t=1}^{T} \left(\frac{t^2}{8} - \frac{t(t-1)}{4} \right) \|\mathbf{w}_t - \mathbf{w}_{t-1}\|_2^2 \right) \leq \frac{8\log n + 2}{T(T+1)},$$

where we used the fact that $\|\mathbf{w}_1\|_2^2 = \|A^\top \mathbf{q}_1\|_2^2 \leq 1$, and $\frac{t^2}{8} \leq \frac{t(t-1)}{4}$ for $t \geq 2$. Based on Theorem 2, we have

$$\min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\overline{\mathbf{w}}_T - \frac{1}{2}\|\overline{\mathbf{w}}_T\|_2^2 = m(\overline{\mathbf{w}}_T) \geq -\frac{8\log n + 2}{T(T+1)} + \max_{\mathbf{w}\in\mathbb{R}^d} m(\mathbf{w}). \tag{35}$$

Next, note that $\overline{\mathbf{w}} = \frac{8}{T(T+1)}\widetilde{\mathbf{w}}_T$. So

$$\begin{aligned}
&\frac{8}{T(T+1)} \min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\widetilde{\mathbf{w}}_T - \frac{64}{2(T(T+1))^2}\|\widetilde{\mathbf{w}}_T\|_2^2 \\
\geq& -\frac{8\log n + 2}{T(T+1)} + \max_{\mathbf{w}\in\mathbb{R}^d} m(\mathbf{w}) \\
\geq& -\frac{8\log n + 2}{T(T+1)} + m\left(\frac{8\|\widetilde{\mathbf{w}}_T\|_2}{T(T+1)} \mathbf{w}^* \right) \\
=& -\frac{8\log n + 2}{T(T+1)} + \frac{8\|\widetilde{\mathbf{w}}_T\|_2}{T(T+1)} \min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{w}^* - \frac{64\|\widetilde{\mathbf{w}}_T\|_2^2}{2(T(T+1))^2}\|\mathbf{w}^*\|_2^2 \\
=& -\frac{8\log n + 2}{T(T+1)} + \frac{8\|\widetilde{\mathbf{w}}_T\|_2\gamma}{T(T+1)} - \frac{64\|\widetilde{\mathbf{w}}_T\|_2^2}{2(T(T+1))^2},
\end{aligned} \tag{36}$$

which implies that

$$\frac{\min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\widetilde{\mathbf{w}}_T}{\|\widetilde{\mathbf{w}}_T\|_2} \geq -\frac{8\log n + 2}{8\|\widetilde{\mathbf{w}}_T\|_2} + \gamma. \tag{37}$$

Finally, note that we always have $\frac{\widetilde{\mathbf{w}}_T^\top \mathbf{w}^*}{\|\widetilde{\mathbf{w}}_T\|_2} \leq 1$, so

$$\begin{aligned}
\|\widetilde{\mathbf{w}}_T\|_2 &\geq \frac{\sum_{t=1}^T \alpha_t}{4} \overline{\mathbf{w}}_T^\top \mathbf{w}^* \\
&= \frac{\sum_{t=1}^T \alpha_t}{4} \left(\frac{1}{\sum_{t=1}^T \alpha_t} \sum_{t=1}^{T} \alpha_t \mathbf{p}_t \right)^\top A\mathbf{w}^* \geq \frac{T(T+1)\gamma}{8}.
\end{aligned} \tag{38}$$

Thus, we have

$$\frac{\min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\mathbf{s}_T}{\|\mathbf{s}_T\|_2} = \frac{\min_{\mathbf{p}\in\Delta^n} \mathbf{p}^\top A\widetilde{\mathbf{w}}_T}{\|\widetilde{\mathbf{w}}_T\|_2} \geq \gamma - \frac{8\log n + 2}{T(T+1)\gamma}.$$

D.2 PROOF OF THEOREM 7

Recall that $\alpha_t = 1$, for all $t \in [T]$, and that the regularizer $\frac{1}{2(q-1)}\|\cdot\|_q^2$ is 1-strongly convex w.r.t. $\|\cdot\|_q$. Then, for the $\mathbf{w}$-player, based on Lemma 3, we have $\forall \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_q \leq 1$,

$$\begin{aligned}
\sum_{t=1}^{T} h_t(\mathbf{w}_t) - h_t(\mathbf{w}) &\leq \frac{1}{2(q-1)\eta^{\mathbf{w}}} + \sum_{t=1}^{T} \frac{\|A^\top(\mathbf{p}_t - \mathbf{p}_{t-1})\|_p^2}{2/\eta^{\mathbf{w}}} \\
&= \frac{1}{2(q-1)\eta^{\mathbf{w}}} + \sum_{t=1}^{T} \frac{\|\sum_{i=1}^{n} y^{(i)}[\mathbf{p}_t - \mathbf{p}_{t-1}]_i \mathbf{x}^{(i)}\|_p^2}{2/\eta^{\mathbf{w}}} \\
&\leq \frac{1}{2(q-1)\eta^{\mathbf{w}}} + \sum_{t=1}^{T} \frac{\left(\sum_{i=1}^{n} |\mathbf{p}_t - \mathbf{p}_{t-1}|_i \|\mathbf{x}^{(i)}\|_p\right)^2}{2/\eta^{\mathbf{w}}} \\
&\leq \frac{1}{2(q-1)\eta^{\mathbf{w}}} + \sum_{t=1}^{T} \frac{\|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2}{2/\eta^{\mathbf{w}}}
\end{aligned} \tag{39}$$

On the other hand, for the $\mathbf{p}$-player, applying Lemma 2, we get

$$\sum_{t=1}^{T} \ell_t(\mathbf{p}_t) - \ell_t(\mathbf{p}) \leq \frac{\log n}{\eta^{\mathbf{p}}} - \sum_{t=1}^{T} \frac{1}{2\eta^{\mathbf{p}}} \|\mathbf{p}_t - \mathbf{p}_{t-1}\|_1^2. \tag{40}$$

Thus, with $\eta^{\mathbf{p}} = 1/\eta^{\mathbf{w}}$, $\eta^{\mathbf{w}} = \sqrt{\frac{1}{2(q-1)\log n}}$, we get

$$m(\overline{\mathbf{w}}_T) \geq \gamma - \frac{\sqrt{\frac{2\log n}{(q-1)}}}{T} = \gamma - \frac{\sqrt{2(p-1)\log n}}{T}, \tag{41}$$

where the equality is based on the relationship between p and q. Hence, we obtain a positive margin when $T > \frac{\sqrt{2(p-1)\log n}}{\gamma}$. On the other hand, let $\alpha \in (0,1)$, and if we would like to obtain

$$\gamma - \frac{\sqrt{2(p-1)\log n}}{T} \geq (1-\alpha)\gamma,$$

we have $T \geq \frac{\sqrt{2(p-1)\log n}}{\alpha\gamma}$.