Technical Report
TTIC-TR-2009-6
September 2009

Convex Games in Banach Spaces

Karthik Sridharan
Toyota Technological Institute at Chicago
karthik@tti-c.org

Ambuj Tewari
Toyota Technological Institute at Chicago
tewari@tti-c.org

ABSTRACT

We study the regret of an online learner playing a multi-round game in a Banach space $\mathfrak{B}$ against an adversary that plays a convex function at each round. We characterize the minimax regret when the adversary plays linear functions in terms of the Rademacher type of the dual of $\mathfrak{B}$. The cases when the adversary plays bounded and uniformly convex functions respectively are also considered. Our results connect online convex learning to the study of the geometry of Banach spaces. We also show that appropriate modifications of the Mirror Descent algorithm from convex optimization can be used to achieve our regret upper bounds. Finally, we provide a version of Mirror Descent that adapts to the changing exponent of uniform convexity of the adversary's functions.

1 Introduction

Online convex optimization [1, 2, 3] has emerged as an abstraction that allows a unified treatment of a variety of online learning problems where the underlying loss function is convex. In this abstraction, a T-round game is played between the learner (or the player) and an adversary. At each round $t \in \{1, \ldots, T\}$, the player makes a move $\mathbf{w}_t$ in some set $\mathcal{W}$. In the learning context, the set $\mathcal{W}$ will represent some hypothesis space. Once the player has made his choice, the adversary then picks a *convex* function ℓ_t from some set $\mathcal{F}$ and the player suffers "loss" $\ell_t(\mathbf{w}_t)$. In the learning context, the adversary's move ℓ_t encodes the data seen at time t and the loss function used to measure the performance of $\mathbf{w}_t$ on that data. As with any abstraction, on one hand, we lose contact with the concrete details of the problem at hand, but on the other hand, we gain the ability to study related problems from a unified point of view. An added benefit of this abstraction is that it connects online learning with geometry of convex sets, theory of optimization and game theory.

An important notion in the online setting is that of the cumulative *regret* incurred by the player which is the difference between the cumulative loss of the player and the cumulative loss for the best fixed move in hindsight. The goal of regret minimizing algorithms is to control the growth rate of the regret as a function of T. There has been a huge amount of work characterizing the best regret rates possible under a variety of assumptions on the player's and adversary's sets $\mathcal{W}$ and $\mathcal{F}$. With a few exceptions that we mention later, most of the work has been in the setting where these sets live in some Euclidean space $\mathbb{R}^d$. Whenever the results do not explicitly involve the dimensionality d, they are also usually applicable in any Hilbert space $\mathfrak{H}$. Our contribution in this paper is to extend the study of optimal regret rates when the set $\mathcal{W}$ lives in a general Banach space $\mathfrak{B}$.

In the Hilbert space setting, it is known that the "degree of convexity" or "curvature" of the functions ℓ_t played by the adversary has a significant impact on the achievable regret rates. For example, if the adversary can play arbitrary convex and Lipschitz functions, the best regret possible is $O(\sqrt{T})$. However, if the adversary is constrained to play *strongly convex* and Lipschitz functions, the regret can be brought down to $O(\log T)$. Further, it is also known, via minimax lower bounds [4], that these are the best possible rates in these situations. In a general Banach space, strongly convex functions might not even exist. We will, therefore, need a generalization of strong convexity called *q-uniform convexity* (strong convexity is 2-uniform convexity). There will, in general, be a number $q^\star \in [2, \infty)$ such that $q^\star$-uniformly convex functions are the "most curved" functions available on $\mathfrak{B}$. There are, again, two extremes: the adversary can play either arbitrary convex-Lipschitz functions or $q^\star$-uniformly convex functions. We show that the minimax optimal rates in these two situations are of the order $\Theta^\star(T^{1/p^\star})$ and $\Theta^\star(T^{2-q^\star})$ respectively[1] where $p^\star$ is the *(Rademacher) type* of the dual $\mathfrak{B}^\star$ of $\mathfrak{B}$. A Hilbert space has $p^\star = q^\star = 2$. We also give upper and lower bounds for the intermediate case when the adversary playes q-uniformly convex functions for $q > q^\star$. This case, as far as we know, has not been analyzed even in the Hilbert space setting.

Another natural game that we have not seen analyzed before is the convex-bounded game: here the adversary plays convex and bounded functions. Of course, being Lipschitz on a bounded domain implies boundedness but the reverse implication is false: a bounded function can have arbitrarily bad Lipschitz constant. For the convex-bounded game, we do not have a tight characterization but we can give non-trivial upper bounds. However, these upper bounds suffice to prove, for example, that the following three properties of $\mathfrak{B}$ are equivalent: (1) the convex-bounded game when the player plays in the unit ball of $\mathfrak{B}$ has non-trivial (i.e. $o(T)$) minimax regret; (2) the corresponding convex-Lipschitz game has non-trivial minimax regret; and (3) the strong law of large numbers holds for bounded i.i.d. random variables in $\mathfrak{B}$.

For most "reasonable" Banach spaces, our results are also constructive. That is, we describe player strategies that achieve the upper bounds we give. These are all based on the *Mirror Descent* algorithm that originated in the convex optimization literature [5]. Usually Mirror Descent is run with a strongly convex function but it turns out that it can also be analyzed in our Banach space setting if it is run with a q-uniformly convex function Ψ. Moreover, with the correct choice of Ψ, it achieves all the upper bounds presented in this paper. Thus, part of our contribution is also to show the remarkable properties of the Mirror Descent algorithm.

[1] Our informal $\Theta^\star(\cdot)$ notation hides factors that are $o(T^\epsilon)$ for every $\epsilon > 0$.

The idea of exploiting minimax-maximin duality to analyze optimal regret rates also appears in the recent work of Abernethy et al. [6]. The earliest papers we know of that explore the connection of the type of a Banach space to learning theory are those of Donahue et al. [7] and Gurvits [8]. More recently, Mendelson and Schechtman [9] gave estimates of the fat-shattering dimension of linear functionals on a Banach space in terms of its type. In the context of online regression with squared loss, Vovk [10] also gives rates worse than $O(\sqrt{T})$ when the class of functions one is competing against is not in a Hilbert space, but in some Banach space.

The rest of the paper is organized as follows. In Section 2, we formally define the minimax and maximin values of the game between a player and an adversary. We also introduce the notions of type, M-type and uniform convexity from functional analysis. Section 3 considers convex-Lipschitz and linear games. A key result in that section is Theorem 3 which gives a characterization of the minimax regret of these games in terms of the type. Convex-bounded games are treated in Section 4 and the equivalence of the strong law of large numbers to the existence of player strategies achieving non-trivial regret guarantees is established (Corollary 5). In Section 5, we describe player strategies based on the Mirror Descent algorithm that achieve the upper bounds presented in the previous two sections. Section 6 consider the case when the adversary plays "curved" functions. Here the regret depends on the exponent of uniform convexity of the functions played by the adversary (Theorem 11). In Section 7, using the techniques of Bartlett et al. [11], we give a player strategy that adapts to the exponent of uniform convexity of the functions being played by the adversary. We conclude in Section 8 by exploring directions for future work including a discussion on how the ideas in this paper might lead to practical algorithms. All proofs omitted from the main body of the paper can be found in the appendix.

2 Preliminaries

2.1 Regret and Minimax Value

Our primary objects of study are certain T-round games where the player $\mathcal{P}$ makes moves in compact convex set $\mathcal{W}$ contained in some (real) separable Banach space $\mathfrak{B}$. The adverary $\mathcal{A}$ plays bounded continuous convex functions on $\mathcal{W}$ chosen from a fixed function class $\mathcal{F}$. We will assume that $\mathcal{F}$ is compact in the topology induced by the norm $\|\ell\| := \sup_{\mathbf{w}\in\mathcal{W}} |\ell(\mathbf{w})|$. The game proceeds as follows.

For $t = 1$ to T

- $\mathcal{P}$ plays $\mathbf{w}_t \in \mathcal{W}$,
- $\mathcal{A}$ plays $\ell_t \in \mathcal{F}$,
- $\mathcal{P}$ suffers $\ell_t(\mathbf{w}_t)$.

For given sequences $\mathbf{w}_{1:T}$, $\ell_{1:T}$, we define the *regret* of $\mathcal{P}$ as,

$$\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) := \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \inf_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \ell_t(\mathbf{w}) \ .$$

Given the tuple $(T, \mathfrak{B}, \mathcal{W}, \mathcal{F})$, we can define the minimax value of the above game as follows.

Definition 1. *Given $T \geq 1$ and $\mathfrak{B}, \mathcal{W}, \mathcal{F}$ satisfying the conditions above, define the minimax value,*

$$V_{T,\mathfrak{B}}(\mathcal{W}, \mathcal{F}) := \inf_{\mathbf{w}_1\in\mathcal{W}} \sup_{\ell_1\in\mathcal{F}} \cdots \inf_{\mathbf{w}_T\in\mathcal{W}} \sup_{\ell_T\in\mathcal{F}} \mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \ .$$

When T and $\mathfrak{B}$ are clear from context, we will simply denote the minimax value by $V(\mathcal{W}, \mathcal{F})$. A *player strategy* (or $\mathcal{P}$-strategy) W is a sequence $(W_1, \ldots, W_T)$ of functions such that $W_t : \mathcal{F}^{t-1} \to \mathcal{W}$. For a strategy W, we define the regret as,

$$\mathrm{Reg}(W, \ell_{1:T}) := \sum_{t=1}^{T} \ell_t(W_t(\ell_{1:t-1})) - \inf_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \ell_t(\mathbf{w}) \ .$$

In terms of player strategies, the minimax value takes a simpler form,

$$V_{T,\mathfrak{B}}(\mathcal{W},\mathcal{F}) = \inf_{W} \sup_{\ell_{1:T}} \mathrm{Reg}(W, \ell_{1:T}) \ ,$$

where the supremum is over all sequences $\ell_{1:T} \in \mathcal{F}^T$. Compactness of $\mathcal{F}$ allows us to define probability measures on $\mathcal{F}^T$. Let Q denote distributions over $\mathcal{F}^T$. We can define the maximin value,

$$U_{T,\mathfrak{B}}(\mathcal{W},\mathcal{F}) := \sup_{Q} \inf_{W} \mathbb{E}_{\ell_{1:T}\sim Q}\left[\mathrm{Reg}(W, \ell_{1:T})\right] \ .$$

By a general minimax theorem [12, Thm. 7.1], under the conditions we have imposed on $\mathcal{W}$ and $\mathcal{F}$, the minimax and maximin values are equal, i.e.

$$V_{T,\mathfrak{B}}(\mathcal{W},\mathcal{F}) = U_{T,\mathfrak{B}}(\mathcal{W},\mathcal{F}) \ . \tag{1}$$

This equality will be the starting point for our subsequent analysis. The equivalence of the minimax and maximin values is not new and can be found in, for instance, [12, p. 31].

2.2 Type, M-type and Uniform Convexity

One of the goals of this paper is to characterize the minimax value in terms of the geometric properties of the space $\mathfrak{B}$ and "degree of convexity" inherent in the functions in $\mathcal{F}$. Among the geometric characteristics of a Banach space $\mathfrak{B}$, the two most useful for us are the *Rademacher type* (or simply *type*) and the *Martingale type* (or *M-type*) of $\mathfrak{B}$. A Banach space $\mathfrak{B}$ *has type* p if there is some constant C such that for any $T \geq 1$ and any $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathfrak{B}$,

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T} \epsilon_t \mathbf{v}_t\right\|\right] \leq C \left(\sum_{t=1}^{T} \|\mathbf{v}_t\|^p\right)^{1/p} \ , \tag{2}$$

where ϵ_t's are i.i.d. Rademacher (symmetric ± 1-valued) random variables. Clearly, any Banach space has type 1. Having type p' for $p' > p$ implies having type p. Therefore, we define

$$p^\star(\mathfrak{B}) := \sup\{p \ : \ \mathfrak{B} \text{ has type } p\} \ . \tag{3}$$

Note that $\mathfrak{B}$ may not have type $p^\star(\mathfrak{B})$, i.e. the supremum above may not be achieved. Also, we always have $1 \leq p^\star(\mathfrak{B}) \leq 2$. A notion related to that of type is that of cotype. A Banach space $\mathfrak{B}$ *has cotype* q if there is some constant C such that for any $T \geq 1$ and any $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathfrak{B}$,

$$\left(\sum_{t=1}^{T} \|\mathbf{v}_t\|^q\right)^{1/q} \leq C\,\mathbb{E}\left[\left\|\sum_{t=1}^{T} \epsilon_t \mathbf{v}_t\right\|\right] \ , \tag{4}$$

where ϵ_t's are i.i.d. Rademacher (symmetric ± 1-valued) random variables.

Beck [13] defined B-convexity to study strong law of large numbers in Banach spaces. A Banach space $\mathfrak{B}$ is *B-convex* if there exists $T > 0$ and an $\epsilon > 0$ such that for any $\mathbf{v}_1, \ldots, \mathbf{v}_T \in \mathfrak{B}$ with $\|\mathbf{v}_t\| \leq 1$, we have

$$\left\|\sum_{t=1}^{T} \epsilon_t \mathbf{v}_t\right\| \leq (1-\epsilon)T \tag{5}$$

for some choices of signs $\epsilon_1, \ldots, \epsilon_T \in \{\pm 1\}$. As it turns out, $\mathfrak{B}$ is B-convex iff $\mathfrak{B}$ has non-trivial type $p > 1$ [14].

A Banach space $\mathfrak{B}$ *has M-type* p if there is some constant C such that for any $T \geq 1$ and martingale difference sequence $\mathbf{d}_1, \ldots, \mathbf{d}_T$ with values in $\mathfrak{B}$,

$$\mathbb{E}\left[\left\|\sum_{t=1}^{T} \mathbf{d}_t\right\|\right] \leq C \left(\sum_{t=1}^{T} \|\mathbf{d}_t\|^p\right)^{1/p} \ . \tag{6}$$

We also define the best M-type possible for a Banach space,

$$p^\star_M(\mathfrak{B}) := \sup \{p \,:\, \mathfrak{B} \text{ has M-type } p\} \ . \tag{7}$$

Similar to the definition of cotype is the definition of M-cotype. A Banach space $\mathfrak{B}$ *has M-cotype* q if there is some constant C such that for any $T \geq 1$ and martingale difference sequence $\mathbf{d}_1, \ldots, \mathbf{d}_T$ with values in $\mathfrak{B}$,

$$\left(\sum_{t=1}^{T} \|\mathbf{d}_t\|^q \right)^{1/q} \leq C \, \mathbb{E}\left[\left\| \sum_{t=1}^{T} \mathbf{d}_t \right\| \right] \ . \tag{8}$$

To measure the "degree of convexity" of the functions played by the adversary, we need the notion of uniform convexity. Let $\|\cdot\|$ be the norm associated with a Banach space $\mathfrak{B}$. A function $\ell : \mathfrak{B} \to \mathbb{R}$ is said to be (C,q)*-uniformly convex* on $\mathfrak{B}$ if there is some constant $C > 0$ such that, for any $\mathbf{v}_1, \mathbf{v}_2 \in \mathfrak{B}$ and any $\theta \in [0,1]$,

$$\ell(\theta \mathbf{v}_1 + (1-\theta)\mathbf{v}_2) \leq \theta \ell(\mathbf{v}_1) + (1-\theta)\ell(\mathbf{v}_2) - \frac{C\theta(1-\theta)}{q} \|\mathbf{v}_1 - \mathbf{v}_2\|^q \ .$$

If $C \geq 1$ we simply say that the function ℓ is q-uniformly convex.

For a convex function $\ell : \mathfrak{B} \to \mathbb{R}$, its *subdifferential* at a point $\mathbf{v}$ is defined as,

$$\partial \ell(\mathbf{v}) = \{ \boldsymbol{\lambda} \in \mathfrak{B}^\star \,:\, \forall \mathbf{v}',\ \ell(\mathbf{v}') \geq \ell(\mathbf{v}) + \boldsymbol{\lambda}(\mathbf{v}' - \mathbf{v}) \} \ ,$$

where $\mathfrak{B}^\star$ denotes the *dual space* of $\mathfrak{B}$. This consists of all continuous linear functions on $\mathfrak{B}$ with norm defined as $\|\ell\|_\star := \sup_{\mathbf{w}:\|\mathbf{w}\|\leq 1} \ell(\mathbf{w})$. If $\partial\ell(\mathbf{v})$ is a singleton then we say ℓ is differentiable at $\mathbf{v}$ and denote the unique member of $\partial\ell(\mathbf{v})$ by $\nabla\ell(\mathbf{v})$. If ℓ is differentiable at $\mathbf{v}_1$, define the *Bregman divergence* associated with ℓ as,

$$\Delta_\ell(\mathbf{v}_1, \mathbf{v}_2) = \ell(\mathbf{v}_1) - \ell(\mathbf{v}_2) - \nabla\ell(\mathbf{v}_1)(\mathbf{v}_1 - \mathbf{v}_2) \ .$$

Recall that a function $\ell : \mathfrak{B} \to \mathbb{R}$ is L-Lipschitz on some set $\mathcal{W} \in \mathfrak{B}$ if for any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{W}$, we have $\ell(\mathbf{v}_1) - \ell(\mathbf{v}_2) \leq L \|\mathbf{v}_1 - \mathbf{v}_2\|$. Given a set $\mathcal{W}$ in a Banach space $\mathfrak{B}$, we define the following natural sets of convex functions on $\mathcal{W}$,

$$\begin{aligned}
\mathrm{lin}(\mathcal{W}) &:= \{\ell \,:\, \ell \text{ is linear and 1-Lipschitz on } \mathcal{W}\} \ , \\
\mathrm{cvx}(\mathcal{W}) &:= \{\ell \,:\, \ell \text{ is convex and 1-Lipschitz on } \mathcal{W}\} \ , \\
\mathrm{bdd}(\mathcal{W}) &:= \{\ell \,:\, \ell \text{ is convex and bounded by 1 on } \mathcal{W}\} \ , \\
\mathrm{cvx}_{q,L}(\mathcal{W}) &= \{\ell \,:\, \ell \text{ is } q\text{-uniformly convex and L-Lipschitz on } \mathcal{W}\} \ .
\end{aligned}$$

In the following sections, we will analyze the minimax value $V(\mathcal{W}, \mathcal{F})$ when the adversary's set of moves is one these 4 sets defined above. For readability, we will drop the dependence of these sets on $\mathcal{W}$ when it is clear from context. For example, we will refer to $V(\mathcal{W}, \mathrm{cvx}(\mathcal{W}))$ simply as $V(\mathcal{W}, \mathrm{cvx})$.

3 Convex-Lipschitz and Linear Games

Given a Banach space $\mathfrak{B}$ with a norm $\|\cdot\|$, denote its unit ball by $U(\mathfrak{B}) := \{\mathbf{v} \in \mathfrak{B} \,:\, \|\mathbf{v}\| \leq 1\}$. Consider the case when the $\mathcal{P}$'s set $\mathcal{W}$ is the unit ball $U(\mathfrak{B})$ for some $\mathfrak{B}$. This setting is not as restrictive as it sounds since any bounded *symmetric* convex set K in a vector space $\mathbf{V}$ gives a Banach space $\mathfrak{B} = (\mathbf{V}, \|\cdot\|_K)$, where we equip $\mathbf{V}$ with the norm,

$$\|\mathbf{v}\|_K := \inf\{\alpha > 0 \,:\, \mathbf{v} \in \alpha K\} \ . \tag{9}$$

Moreover, (the closure of) K is the unit ball of this Banach space.

So, fix $\mathfrak{B}$ and consider the case $\mathcal{W} = U(\mathfrak{B})$, $\mathcal{F} = \mathrm{cvx}(\mathcal{W})$. Theorem 14 given in [4] gives us $V(\mathcal{W}, \mathrm{cvx}) = V(\mathcal{W}, \mathrm{lin})$. We are therefore led to consider the case $\mathcal{W} = U(\mathfrak{B})$, $\mathcal{F} = \mathrm{lin}(\mathcal{W})$. Note that $\mathrm{lin}(\mathcal{W})$ is simply the unit ball $U(\mathfrak{B}^\star)$. The theorem below relates the minimax value $V(U(\mathfrak{B}), \mathrm{lin})$ to the behaviour of random walks in $\mathfrak{B}^*$ generated by Rademacher random variables.

Theorem 2. *Let $\epsilon_1, \ldots, \epsilon_T$ be i.i.d. Rademacher random variables. Define*

$$\mathcal{R}(\mathfrak{B}) := \sup_{\ell_{1:T}} \mathbb{E}\left[\left\|\sum_{t=1}^{T} \epsilon_t \ell_t\right\|_\star\right] ,$$

where the supremum is over all sequences $\ell_{1:T}$ such that $\ell_t \in U(\mathfrak{B}^\star)$. Then, the minimax value $V(U(\mathfrak{B}), \mathrm{lin})$ is bounded as,

$$\mathcal{R}(\mathfrak{B}) \leq V(U(\mathfrak{B}), \mathrm{lin}) \leq 2\mathcal{R}(\mathfrak{B}) .$$

Proof. Note that when $\mathcal{W} = U(\mathfrak{B})$ and $\mathcal{F} = \mathrm{lin}(\mathcal{W})$, all ℓ_t's are linear functions. Therefore,

$$\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \epsilon_t \ell_t(\mathbf{w}) = \left\|\sum_{t=1}^{T} \epsilon_t \ell_t\right\|_\star .$$

The theorem then follows by combining Lemmas 14 and 15. □

Given the above result, we can now characterize the minimax value $V(U(\mathfrak{B}), \mathrm{lin})$ in terms of the $p^\star(\mathfrak{B}^\star)$ where $\mathfrak{B}^\star$ is the dual space of $\mathfrak{B}$.

Theorem 3. *For all $p \in [1, p^\star(\mathfrak{B}^\star))$, there exists a constant C such that,*

$$T^{1/p^\star(\mathfrak{B}^\star)} \leq V(U(\mathfrak{B}), \mathrm{lin}) = V(U(\mathfrak{B}), \mathrm{cvx}) \leq CT^{1/p} . \tag{10}$$

If the supremum in (3) *is achieved, the upper bound also holds for $p = p^\star(\mathfrak{B}^\star)$.*

Proof. The upper bound follows from Theorem 2 and the definition of type (2). For the lower bound, note that any finite dimensional Banach space has $p^\star(\mathfrak{B}^\star) = 2$ with a possibly dimension dependent constant. In this case, the lower bound of $\sqrt{T}$ for the linear game game is easy: the adversary picks some non-zero vector, say $\ell \in U(\mathfrak{B}^\star)$, and plays ℓ or $-\ell$ at random. Therefore, assume $\mathfrak{B}$ is infinite dimensional. By Theorem 3.5 of [15] (see also Remark 3.4), we see that there exists $\ell_1, ..., \ell_T \in U(\mathfrak{B}^\star)$ such that for any choice of $\epsilon_1, ..., \epsilon_T$,

$$\left\|\sum_{t=1}^{T} \epsilon_t \ell_t\right\|_\star \geq \left(\sum_{t=1}^{T} |\epsilon_t|^{p^\star(\mathfrak{B}^\star)}\right)^{\frac{1}{p^\star(\mathfrak{B}^\star)}} = T^{1/p^\star(\mathfrak{B}^\star)} .$$

Hence, by Theorem 2, we get the lower bound. □

Although we have stated the above theorem for the special case when $\mathcal{W} = U(\mathfrak{B})$ and $\mathcal{F} = \mathrm{lin}(U(\mathfrak{B}))$, it actually gives us regret rates when $\mathcal{P}$ plays in $r\,U(\mathfrak{B})$ and $\mathcal{A}$ plays L-Lipschitz linear functions via the following equality which is easy to prove from first principles,

$$V(r\,U(\mathfrak{B}), L\ \mathrm{lin}(U(\mathfrak{B}))) = r \cdot L \cdot V(U(\mathfrak{B}), \mathrm{lin}(U(\mathfrak{B}))) . \tag{11}$$

4 Convex-Bounded Games

Another natural game we consider is one in which $\mathcal{P}$ plays from some convex set $\mathcal{W}$ and $\mathcal{A}$ plays some convex function *bounded* by 1. In the following theorem, we bound the value of such a game.

Theorem 4. *For all $p \in [1, p^\star(\mathfrak{B}^\star))$, there is a constant C such that,*

$$T^{1/p^\star(\mathfrak{B}^\star)} \leq V(U(\mathfrak{B}), \mathrm{bdd}) \leq CT^{1/p+1/2q} . \tag{12}$$

where $q = \frac{p}{p-1}$. If the supremum in (3) *is achieved, the upper bound also holds for $p = p^\star(\mathfrak{B}^\star)$.*

Proof of Theorem 4. Let us actually consider the case $\mathcal{W} = r\,U(\mathfrak{B})$ and $\mathcal{F} = \mathrm{bdd}(r\,U(\mathfrak{B}))$. The bounds will turn out to be independent of r. Note that we have the inclusion,

$$\mathrm{bdd}(r\,U(\mathfrak{B})) \supseteq \frac{1}{r}\,\mathrm{lin}(U(\mathfrak{B}))$$

which implies

$$V(r\,U(\mathfrak{B}), \mathrm{bdd}(r\,U(\mathfrak{B}))) \geq V\left(r\,U(\mathfrak{B}), \frac{1}{r}\,\mathrm{lin}(U(\mathfrak{B}))\right) .$$

The lower bound is now immediate due to lower bound on linear game on unit ball (Theorem 3) and property (11).

For the upper bound, note that any convex function bounded by 1 on the scaled ball $r\,U(\mathfrak{B})$ is $\frac{2}{\epsilon r}$ Lipschitz on the ball of radius $(1-\epsilon)r$ [21]. Hence, by upper bound in Theorem 3 and property (11), we see that there exists a strategy say W whose regret on the ball of radius $r(1-\epsilon)$ is bounded by $\frac{C}{\epsilon}T^{\frac{1}{p}}$ for any $p \in [1, p^\star(\mathfrak{B}^\star)$. That is

$$\sum_{t=1}^{T} \ell_t(W_t) - \operatorname*{argmin}_{\mathbf{w}\in r(1-\epsilon)U(\mathfrak{B})} \sum_{t=1}^{T} \ell_t(\mathbf{w}) \leq \frac{C}{\epsilon}T^{1/p}, \quad \forall p \in [1, p^\star(\mathfrak{B}^\star)) \tag{13}$$

Let $\mathbf{w}^\star = \operatorname*{argmin}_{\mathbf{w}\in r\,U(\mathfrak{B})} \sum_{t=1}^{T} \ell_t(\mathbf{w})$. Now we consider two cases, first when $\mathbf{w}^\star \in (1-\epsilon)r\,U(\mathfrak{B})$ In this case the regret of the strategy on the unit ball is bounded by $\frac{C}{\epsilon}T^{1/p}$ for all $p \in [1, p^\star(\mathfrak{B}^\star))$. On the other hand if $\mathbf{w}^\star \notin (1-\epsilon)r\,U(\mathfrak{B})$, then define $\mathbf{w}^\star_\epsilon = \frac{r(1-\epsilon)\mathbf{w}^\star}{\|\mathbf{w}^\star\|}$. Note that $\mathbf{w}^\star_\epsilon \in r(1-\epsilon)U(\mathfrak{B})$. In this case by convexity of $\sum_{t=1}^{T} \ell_t(\mathbf{w})$, we have that

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(\mathbf{w}^\star_\epsilon) &\leq \frac{r(1-\epsilon)}{\|\mathbf{w}^\star\|} \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) + \left(1 - \frac{r(1-\epsilon)}{\|\mathbf{w}^\star\|}\right) \sum_{t=1}^{T} \ell_t(0) \\
&\leq \frac{r(1-\epsilon)}{\|\mathbf{w}^\star\|} \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) + \left(1 - \frac{r(1-\epsilon)}{\|\mathbf{w}^\star\|}\right) T
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(\mathbf{w}^\star_\epsilon) - \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) &\leq \left(\frac{r(1-\epsilon)}{\|\mathbf{w}^\star\|} - 1\right) \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) + \left(1 - \frac{r(1-\epsilon)}{\|\mathbf{w}^\star\|}\right) T \\
&\leq 2\left(1 - \frac{r(1-\epsilon)}{\|\mathbf{w}^\star\|}\right) T
\end{aligned}$$

However, since $\|\mathbf{w}^\star\| \leq r$ we see that

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}^\star_\epsilon) - \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) \leq 2\epsilon T$$

Combining with (13) we see that for any $p \in [1, p^\star(\mathfrak{B}^\star))$,

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{w}^\star) \leq \frac{C}{\epsilon}T^{1/p} + 2\epsilon T$$

Choosing $\epsilon = \sqrt{\frac{C}{2T^{1/q}}}$ we get that in either of the cases (ie. $\|\mathbf{w}^\star\| \leq r(1-\epsilon)$ and otherwise), we have that for any $p \in [1, p^\star(\mathfrak{B}^\star))$

$$\sum_{t=1}^{T} \ell_t(W_t) - \operatorname*{argmin}_{\mathbf{w}\in U(\mathfrak{B})} \sum_{t=1}^{T} \ell_t(\mathbf{w}) \leq \sqrt{2C}\,T^{\frac{1}{p}+\frac{1}{2q}}$$

□

Although we have stated the above result for the unit ball, the proof given above shows that the bounds are independent of the radius of the ball in which the player is playing.

Theorems 3 and 4 imply the following interesting corollary.

Corollary 5. *The following statements are equivalent :*

1. $V(U(\mathfrak{B}), \mathrm{bdd}) = o(T)$.
2. $V(U(\mathfrak{B}), \mathrm{cvx}) = o(T)$.
3. $\mathfrak{B}^\star$ *has non-trivial type*
4. *Both* $\mathfrak{B}$ *and* $\mathfrak{B}^\star$ *are B-convex.*
5. *Law of large numbers for i.i.d. bounded random variables holds in both* $\mathfrak{B}$ *and* $\mathfrak{B}^\star$.

Proof of Corollary 5. The implications **3** $\Rightarrow$ **1** and **3** $\Rightarrow$ **2** follow from the upper bounds in Theorems 3 and 4. The reverse implications **1** $\Rightarrow$ **3** and **2** $\Rightarrow$ **3**, in turn, follow from the lower bounds in those theorems. The equivalence of **3** and **4** is due to deep results of Pisier [14]. The equivalence of **4** and **5** is due to Beck [13]. □

The convex-Lipschitz games (and q-uniformly convex-Lipschitz games considered below) depend, by definition, not only on the player's set $\mathcal{W}$ but also on the norm $\|\cdot\|$ of the underlying Banach space $\mathfrak{B}$. This is because $\mathcal{A}$'s functions are required to be Lipschitz w.r.t. $\|\cdot\|$. However, note that the convex-bounded game can be defined only in terms of the player set $\mathcal{W}$. Hence, one would expect the value of the game to be characterized solely by properties of set $\mathcal{W}$. This is what the following corollary confirms.

Corollary 6. *Let* $\mathcal{W}$ *be any symmetric bounded convex subset of a vector space* $\mathbf{V}$. *The value of the bounded convex game on* $\mathcal{W}$ *is non-trivial (i.e.* $o(T)$*) iff the Banach space* $(\mathbf{V}, \|\cdot\|_{\mathcal{W}})$ *(where* $\|\cdot\|_{\mathcal{W}}$ *is defined as in* (9)*) is B-convex.*

5 Strategy for the Player

The upper bound on the regret given in Theorem 3 relies on Theorem 2 which, in turn, uses minimax-maximin equality. The proof, therefore, is not constructive and does not yield a strategy for $\mathcal{P}$ achieving the upper bound. In this section, we provide a strategy whose regret achieves the minimax value for a wide class of Banach spaces. The strategy we consider is known as Mirror Descent and is given as Algorithm 1 below. The following proposition gives a regret bound for Mirror Descent.

Algorithm 1 Mirror Descent (Parameters : $\eta > 0$, $\Psi : \mathfrak{B} \to \mathbb{R}$ which is uniformly convex)

```
for t = 1 to T do
  Play w_t and receive ℓ_t
  w'_{t+1} ← ∇Ψ*(∇Ψ(w_t) − ηλ_t) where λ_t ∈ ∂ℓ_t(w_t)
  Update w_{t+1} ← argmin_{w∈W} Δ_Ψ(w, w'_{t+1})
end for
```

Proposition 7. *Suppose* $\mathcal{W} \subseteq \mathfrak{B}$ *is such that* $\|\mathbf{w}\| \le B$. *Let MD denote the* $\mathcal{P}$*-strategy obtained by running Mirror Descent with a function* Ψ *that is q-uniformly convex on* $\mathfrak{B}$ *and C-Lipschitz on* $\mathcal{W}$, *and the learning rate* $\eta = (BC/T)^{1/p} \cdot (1/L)$. *Here,* $p = q/(q-1)$ *is the dual exponent of q. Then, for all sequences* $\ell_{1:T}$ *such that* ℓ_t *is L-Lipschitz on* $\mathcal{W}$, *we have,*

$$\mathrm{Reg}(MD, \ell_{1:T}) = O\left((BC)^{1/q} \cdot L \cdot T^{1/p}\right)$$

Proof. For $\boldsymbol{\lambda} \in \mathfrak{B}^\star, \mathbf{w} \in \mathfrak{B}$ we denote the pairing $\boldsymbol{\lambda}(\mathbf{w})$ by $\langle \boldsymbol{\lambda}, \mathbf{w} \rangle$ where $\langle \cdot, \cdot \rangle : \mathfrak{B}^\star \times \mathfrak{B} \to \mathbb{R}$. This pairing is bilinear but not symmetric. We will first show that, for any $\mathbf{w} \in \mathcal{W}$,

$$\eta \langle \boldsymbol{\lambda}_t, \mathbf{w}_t - \mathbf{w} \rangle \le \Delta_\Psi(\mathbf{w}, \mathbf{w}_t) - \Delta_\Psi(\mathbf{w}, \mathbf{w}_{t+1}) + \frac{\eta^p}{p} \|\boldsymbol{\lambda}_t\|_\star^2 , \tag{14}$$

where $p = q/(q-1)$. We have,

$$\begin{aligned}
\eta \langle \boldsymbol{\lambda}_t, \mathbf{w}_t - \mathbf{w} \rangle &= \langle \eta \boldsymbol{\lambda}_t, \mathbf{w}_t - \mathbf{w}_{t+1} + \mathbf{w}_{t+1} - \mathbf{w} \rangle \\
&= \langle \eta \boldsymbol{\lambda}_t, \mathbf{w}_t - \mathbf{w}_{t+1} \rangle + \langle \eta \boldsymbol{\lambda}_t, \mathbf{w}_{t+1} - \mathbf{w} \rangle \\
&= \underbrace{\langle \eta \boldsymbol{\lambda}_t, \mathbf{w}_t - \mathbf{w}_{t+1} \rangle}_{s_1} + \underbrace{\langle \eta \boldsymbol{\lambda}_t + \nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w} \rangle}_{s_2} \\
&\quad + \underbrace{\langle \nabla\Psi(\mathbf{w}_t) - \nabla\Psi(\mathbf{w}_{t+1}), \mathbf{w}_{t+1} - \mathbf{w} \rangle}_{s_3}
\end{aligned} \tag{15}$$

Now, by definition of the dual norm and the fact that $ab \le \frac{a^p}{p} + \frac{b^q}{q}$ for any $a, b \ge 0$, we get

$$s_1 \le \|(\mathbf{w}_t - \mathbf{w}_{t+1})\| \cdot \|\eta \boldsymbol{\lambda}\|_\star \le \frac{1}{q}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^q + \frac{1}{p}\|\eta \boldsymbol{\lambda}_t\|_*^p .$$

By the definition of the update, $\mathbf{w}_{t+1}$ minimizes

$$\langle \eta \boldsymbol{\lambda}_t - \nabla\Psi(\mathbf{w}_t), \mathbf{w} \rangle + \Psi(\mathbf{w})$$

over $\mathbf{w} \in \mathcal{W}$. Therefore, $s_2 \le 0$. Using simple algebraic manipulations, we get

$$s_3 = \Delta_\Psi(\mathbf{w}, \mathbf{w}_t) - \Delta_\Psi(\mathbf{w}, \mathbf{w}_{t+1}) - \Delta_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t) \ .$$

Plugging this into (15), we get

$$\begin{aligned}
\eta \langle \boldsymbol{\lambda}_t, \mathbf{w}_t - \mathbf{w} \rangle &\le \Delta_\Psi(\mathbf{w}, \mathbf{w}_t) - \Delta_\Psi(\mathbf{w}, \mathbf{w}_{t+1}) + \frac{\eta^p}{p}\|\boldsymbol{\lambda}_t\|_\star^2 \\
&\quad + \underbrace{\frac{1}{q}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^q - \Delta_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t)}_{s_4}
\end{aligned}$$

Using that Ψ is q-uniformly convex on $\mathfrak{B}$ implies that $s_4 \le 0$. So, we get (14).

We can now bound the regret as follows. For any $\mathbf{w} \in \mathcal{W}$, since $\boldsymbol{\lambda}_t \in \partial \ell_t(\mathbf{w}_t)$, we have,

$$\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}) \le \langle \boldsymbol{\lambda}_t, \mathbf{w}_t - \mathbf{w} \rangle \ .$$

Combining this with (14) and summing over t gives,

$$\sum_{t=1}^{T} (\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w})) \le \frac{\Delta_\Psi(\mathbf{w}, \mathbf{w}_1) - \Delta_\Psi(\mathbf{w}, \mathbf{w}_{T+1})}{\eta} + \frac{\eta^{p-1}}{p} \sum_{t=1}^{T} \|\boldsymbol{\lambda}_t\|_\star^p .$$

Now, $\Delta_\Psi(\mathbf{w}, \mathbf{w}_{T+1}) \ge 0$ and $\Delta_\Psi(\mathbf{w}, \mathbf{w}_1) \le 2BC$. Further $\|\boldsymbol{\lambda}_t\|_\star \le L$ since ℓ_t is L-Lipschitz. Plugging these above and optimizing over η gives,

$$\sum_{t=1}^{T} (\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w})) \le O\left((BC)^{1/q} \cdot L \cdot T^{1/p} \right) \ .$$

□

This gives us hope that we can achieve the upper bound in (10) if we can find a $\frac{p}{p-1}$-uniformly convex function on a given Banach space for any $p \in [1, p^\star(\mathfrak{B}^\star))$. Unfortunately, such functions might not exist for all p in this range. However, the following remarkable theorem of Pisier [16] guarantees that such functions do exist for $p \in [1, p_M^\star(\mathfrak{B}^\star))$.

Theorem (Pisier). *A Banach space $\mathfrak{B}$ has M-cotype q iff there exists a q-uniformly convex function on $\mathfrak{B}$.*

Now consider some $p \in [1, p_M^\star(\mathfrak{B}^\star))$. Then, by definition, $\mathfrak{B}^\star$ has M-type p. It is a fact that $\mathfrak{B}^\star$ has M-type p iff $\mathfrak{B}$ has M-cotype $\frac{p}{p-1}$ [15, Chapter 6]. Thus, $\mathfrak{B}$ has M-cotype $\frac{p}{p-1}$. Now, Pisier's theorem gives us a $\frac{p}{p-1}$-uniformly convex function on $\mathfrak{B}$.

Pisier's theorem addresses the issue of the non-constructive nature of the upper bound in (10) to a large extent. In fact, if $p^\star(\mathfrak{B}^\star) = p_M^\star(\mathfrak{B}^\star)$ then the upper bound is actually achieved by using Mirror Descent with an appropriate Ψ. If $p^\star(\mathfrak{B}^\star) > p_M^\star(\mathfrak{B}^\star)$ then we do not know how to achieve the upper bound for p's in the range $(p_M^\star(\mathfrak{B}^\star), p^\star(\mathfrak{B}^\star))$. However, this only happens quite rarely. In particular, if $\mathfrak{B}$ has local unconditional structure and $\mathfrak{B}^\star$ has type p for some $p > 1$ then it also has M-type p [16]. Most "reasonable" spaces have local unconditional structure[2]. Thus, for the remainder of this paper, we make the following assumption.

Assumption 8. *The Banach space $\mathfrak{B}$ has local unconditional structure [18, p. 59].*

Under this mild assumption, we can show that Mirror Descent achieves the upper bound in (10).

Theorem 9. *For each $p \in [1, p^\star(\mathfrak{B}^\star))$, there exists a Ψ such that the upper bound (10) is achieved by Mirror Descent using this Ψ. If the supremum in (3) is achieved, then p can also be $p^\star(\mathfrak{B}^\star)$.*

Proof. First, if $p^\star(\mathfrak{B}^\star) = 1$ then the upper bound is trivial. On the other hand, when type is non-trivial, then type p implies cotype $q = \frac{p}{p-1}$. Under Assumption 8, we have that cotype q implies M-cotype q. Therefore, for each $p \in (1, p^\star(\mathfrak{B}^\star))$, $\mathfrak{B}$ is of M-cotype q. By Theorem (Pisier), there exists a q-uniformly convex function on $\mathfrak{B}$. Using this function in the Mirror Descent algorithm, Proposition 7 yields the required upper bound. ☐

The below corollary shows that the Mirror Descent can also be used to play the convex-bounded game and that the upper bound in Theorem 4 can be achieved.

Corollary 10. *For each $p \in [1, p^\star(\mathfrak{B}^\star))$, there exists a Ψ and a choice of $\epsilon > 0$ such that if we use Mirror Descent with this Ψ, restricting the player's set to the ball of radius $(1-\epsilon)$, then the upper bound in (12) is achieved by this strategy.*

Proof. The Corollary follows immediately from proof of Theorem 4 where we can replace the strategy we used on the $1-\epsilon$ ball to the MD algorithm on that ball. ☐

6 Uniformly Convex-Lipschitz Games

For any Hilbert space $\mathfrak{H}$, it is known that $V(U(\mathfrak{H}), \mathrm{cvx}_{2,L})$ is much smaller than $V(U(\mathfrak{H}), \mathrm{lin})$, i.e. the game is much easier for $\mathcal{P}$ if $\mathcal{A}$ plays 2-uniformly convex (also called *strongly convex*) functions. In fact, it is known that $V(U(\mathfrak{H}), \mathrm{cvx}_{2,L}) = \Theta(L^2 \log T)$ while $V(U(\mathfrak{H}), \mathrm{lin}) = \Theta(\sqrt{T})$. This suggests that we should get a rate between $\log(T)$ and $\sqrt{T}$ if $\mathcal{A}$ plays q-uniformly convex functions in a Hilbert space $\mathfrak{H}$ for some $q > 2$. As far as we know, there is no characterization of the achievable rates for these intermediate situations even for Hilbert spaces. Our next result provides upper and lower bounds for $V(U(\mathfrak{B}), \mathrm{cvx}_{q,L})$ in a Banach space, when the exponent of $\mathcal{A}$'s uniform convexity lies in an intermediate range between its minimum possible value $q^\star$ and its maximum value ∞. It is easy to see that under Assumption 8, the minimum possible value $q^\star$ is $p^\star(\mathfrak{B}^\star)/(p^\star(\mathfrak{B}^\star)-1)$.

Theorem 11. *Let $q^\star = \frac{p^\star(\mathfrak{B}^\star)}{p^\star(\mathfrak{B}^\star)-1}$ and $q \in (q^\star, \infty)$. Let $p = q/(q-1)$ be the dual exponent of q. Then, as long as $\mathrm{cvx}_{q,L}$ is non-empty, there exists K that depends on L such that,*

$$\left(1 - \frac{1}{L}\right) \frac{1}{p} T^{1-p+\frac{p}{p^\star(\mathfrak{B}^\star)}} \le V(U(\mathfrak{B}), \mathrm{cvx}_{q,L}) \le K T^{\min\{2-p, 1/p^\star(\mathfrak{B}^\star)\}} . \tag{16}$$

[2] In fact, constructing a Banach space without local unconditional structure took some effort [17].

Proof. We start by proving the lower bound. To this end note that if $\mathrm{cvx}_{q,L}$ is non-empty, then the adversary plays L-Lipschitz, q-uniformly convex loss functions. Note that given such a function, there exists a norm $|\cdot|$ such that $|\cdot| \le \|\cdot\| \le L|\cdot|$ (ie. an equivalent norm) and $\frac{1}{q}|\cdot|^q$ is a q-uniformly convex function [22]. Given this we can consider a game where adversary plays only functions from the set

$$\mathrm{lincvx}_{q,L}(\mathcal{W}) := \{\ell(\mathbf{w}) = \langle \mathbf{w}, \mathbf{x}\rangle + \frac{1}{q}|\mathbf{w}|^q : |x|_\star \le L-1\}$$

Now first of all note that since the above is L-Lipschitz w.r.t. $|\cdot|$, it is automatically L-Lipchitz w.r.t. $\|\cdot\|$. Hence $\mathrm{lincvx}_{q,L} \subseteq \mathrm{cvx}_{q,L}$, and so we have that $V(U(\mathfrak{B}), \mathrm{lincvx}_{q,L}) \le V(U(\mathfrak{B}), \mathrm{cvx}_{q,L})$ However note that

$$V(U(\mathfrak{B}), \mathrm{lincvx}_{q,L}) = \inf_W \sup_P \mathbb{E}_{\ell_{1:T}\sim P} \mathrm{Reg}(W, \ell_{1:T}) \tag{17}$$

Note that

$$\begin{aligned}
\mathrm{Reg}(W, \ell_{1:T}) &= \sum_{t=1}^T \left(\langle \mathbf{x}_t, \mathbf{w}_t\rangle + \frac{|\mathbf{w}_t|^q}{q}\right) - \inf_{\mathbf{w}\in U(\mathfrak{B})} \sum_{t=1}^T \left(\langle \mathbf{x}_t, \mathbf{w}\rangle + \frac{|\mathbf{w}|^q}{q}\right) \\
&= \sum_{t=1}^T \left(\langle \mathbf{x}_t, \mathbf{w}_t\rangle + \frac{|\mathbf{w}_t|^q}{q}\right) + T \sup_{\mathbf{w}\in U(\mathfrak{B})} \left(\left\langle -\frac{1}{T}\sum_{t=1}^T \mathbf{x}_t, \mathbf{w}\right\rangle - \frac{|\mathbf{w}|^q}{q}\right) \\
&\ge \sum_{t=1}^T \left(\langle \mathbf{x}_t, \mathbf{w}_t\rangle + \frac{\|\mathbf{w}_t\|^q}{Lq}\right) + T \sup_{\mathbf{w}\in U(\mathfrak{B})} \left(\left\langle -\frac{1}{T}\sum_{t=1}^T \mathbf{x}_t, \mathbf{w}\right\rangle - \frac{\|\mathbf{w}\|^q}{q}\right) \\
&= \sum_{t=1}^T \left(\langle \mathbf{x}_t, \mathbf{w}_t\rangle + \frac{|\mathbf{w}_t|^q}{Lq}\right) + \frac{T\left\|-\frac{1}{T}\sum_{t=1}^T \mathbf{x}_t\right\|_\star^p}{p}
\end{aligned}$$

Where the last step is by definition of convex dual of $\frac{\|\cdot\|^q}{q}$. Now note that since we have a supremum over distribution in (17), and so we can lower bound the value by picking the particular distribution where each $\mathbf{x}_t$ is got by randomly sampling rademacher variable ϵ_t and setting $\mathbf{x}_t = \epsilon_t \mathbf{y}_t$ where $\mathbf{y}_1, ..., \mathbf{y}_T \in \mathfrak{B}^\star$ are fixed (to be set later). Thus we see that for any sequence $\mathbf{y}_1, ..., \mathbf{y}_T$, we have that

$$\begin{aligned}
V(U(\mathfrak{B}), \mathrm{lincvx}_{q,L}) &\ge \inf_W \mathbb{E}\left[\sum_{t=1}^T \left(\langle \epsilon_t \mathbf{y}_t, \mathbf{w}_t\rangle + \frac{\|\mathbf{w}_t\|^q}{Lq}\right) + \frac{T^{1-p}\left\|-\sum_{t=1}^T \epsilon_t \mathbf{y}_t\right\|_\star^p}{p}\right] \\
&= \inf_W \mathbb{E}\left[\sum_{t=1}^T \frac{\|\mathbf{w}_t\|^q}{Lq}\right] + \frac{T^{1-p}\mathbb{E}\left[\left\|-\sum_{t=1}^T \epsilon_t \mathbf{y}_t\right\|_\star^p\right]}{p} \\
&= \frac{T^{1-p}\mathbb{E}\left[\left\|-\sum_{t=1}^T \epsilon_t \mathbf{y}_t\right\|_\star^p\right]}{p} \ge \frac{T^{1-p}\left(\mathbb{E}\left[\left\|-\sum_{t=1}^T \epsilon_t \mathbf{y}_t\right\|_\star\right]\right)^p}{p}
\end{aligned} \tag{18}$$

where the first equality is because $\mathbf{w}_t$ is only dependent on the history and so the expectation over each $\epsilon_t \mathbf{y}_t$ is 0 and the last step is due to Jensen's inequality. However since $\ell_{p^\star(\mathfrak{B}^\star)}$ is uniformly isomorphic to $\mathfrak{B}^\star$ [15], we see that there exists $\mathbf{y}_1, ..., \mathbf{y}_T \in \frac{L-1}{L} U(\mathfrak{B}^\star)$ such that for any $\epsilon_1, ..., \epsilon_T \in \{-1, 1\}$, we have that

$$\left\|-\sum_{t=1}^T \epsilon_t \mathbf{y}_t\right\|_\star \ge \frac{L-1}{L} T^{\frac{1}{p^\star(\mathfrak{B}^\star)}}$$

However since $\frac{1}{L}|\cdot|_\star \le \|\cdot\|_\star \le |\cdot|_\star$, we see that if $\|\mathbf{y}_t\|_\star \le \frac{L-1}{L}$ then, $|\mathbf{y}_t|_\star \le L-1$ and so using these $\mathbf{y}_1, ..., \mathbf{y}_T$ in (18) we conclude that

$$V(U(\mathfrak{B}), \mathrm{lincvx}_{q,L}) \ge \left(1 - \frac{1}{L}\right) \frac{T^{1-p+\frac{p}{p^\star(\mathfrak{B}^\star)}}}{p} \tag{19}$$

Note that the above lower bound becomes 0 when $L = 1$ but however in that case it means that the adversary is forced to play 1-Lipschitz, q-uniformly convex function. However since from each q-uniformly convex L-Lipschitz convex function we can build an equivalent norm with distortion $1/L$, this means that the function the adversary plays can be used to construct the original norm itself. From the construction in [22] it becomes clear that the functions the adversary can play can be merely the norm plus some constant and so the lower bound of 0 is real.

Now we turn to proving the upper bound, for this we simply appeal to Theorem 12 to give a constructive proof for the theorem. Consider the regret of the mirror descent algorithm when we run it using a $q^\star$-uniformly convex function Ψ that is C-Lipschitz on the unit ball. Here, for simplicity, we assume that the supremum is achieved in (3) (otherwise we can pick a Ψ that is q'-uniformly convex for $q' = p'/(p'-1)$ where $p' = p^\star(\mathfrak{B}^\star) - 1/\log T$ and pay a constant factor in the upper bound). Note that in the case when $q > q^\star + 1$, we have that each $\sigma_t^\star = 0$ and so by Theorem 12 we have that,

$$\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \le 2 \min_{\lambda_1,\dots,\lambda_T} \sum_{t=1}^{T} \frac{(L+C)^{p^*}}{\left(\sum_{j\le t} \lambda_j\right)^{p^*-1}} + 2\sum_{t=1}^{T} \lambda_t C \le 2 \min_{\lambda} \frac{(L+C)^{p^*} T^{2-p^*}}{\lambda^{p^*-1}} + 2T\lambda C$$

Using $\lambda = \frac{L+C}{T^{1/q^\star}(2C)^{1/p^\star}}$ we see that

$$\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \le 8(2C)^{1/q^\star}(L+C)T^{1/p^\star}$$

On the other hand when $q \le q^\star + 1$, using the upper bound in the theorem with $\lambda_t = 0$ for all t we see that since all $q_t = q$ and all $\sigma_t^\star = 1$ and $L_t = L$ we find that the regret of the adaptive algorithm is bounded as

$$\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \le \sum_{t=1}^{T} \left(\frac{2(L+C)^p}{t^{p-1}} + \frac{2(L+C)^{p^\star}}{t^{p^\star-1}} \right)$$

Hence we in fact get the bound,

$$\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \le \sum_{t=1}^{T} \left(\frac{2(L+C)^p}{t^{p-1}} + \frac{2(L+C)^{p^\star}}{t^{p^\star-1}} \right) \le \sum_{t=1}^{T} \frac{4(L+C)^p}{t^{p-1}} \le 4(L+C)^p \int_1^T \frac{1}{t^{p-1}} dt$$

Hence we see that for $p < p^\star(\mathfrak{B}^\star) \le 2$, $\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \le \frac{4(L+C)^p}{2-p} T^{2-p}$. Since the regret of the adaptive algorithm bounds the value of the game, we see that by picking constant $K = \max\{\frac{4(L+C)^p}{2-p}, 8(2C)^{1/q^\star}(L+C)\}$ we get that :

$$V\left(U\left(\mathfrak{B}\right), \mathrm{cvx}_{q,L}\right) \le K T^{\min\{2-p, 1/p^\star(\mathfrak{B}^\star)\}} \tag{20}$$

Now combining Equations 19 and 20, we get the theorem statement. ☐

The upper and lower bounds do not match in general and it is an interesting open problem to remove this gap. Note that the upper and lower bounds do match for the two extreme cases $q \to q^\star$ and $q \to \infty$. When $q \to q^\star$, then both lower and upper bound exponents tend to $2 - p^\star(\mathfrak{B}^\star)$. On the other hand, when $q \to \infty$, both exponents tend to $1/p^\star(\mathfrak{B}^\star)$.

7 Adaptive Player Strategy

In the previous section, we analyzed the game where $\mathcal{A}$ plays q-uniformly convex Lipschitz functions. A natural extension is a game where at round t, $\mathcal{A}$ plays q_t uniformly convex functions. In this section, we give an adaptive player strategy for such games that achieves the upper bound in Theorem 11 whenever the adversary plays only q-uniformly convex functions on all rounds and in general gets intermediate rates when the modulus of convexity on each round is different.

Algorithm 2 Adaptive Mirror Descent (Parameters : $\Psi : \mathfrak{B} \to \mathbb{R}$ which is $q^\star$-uniformly convex)

$C \leftarrow$ Lipschitz constant of Ψ on $U(\mathfrak{B})$, $\mathbf{w}_1 \leftarrow \mathbf{0}$, $\Phi_1 \leftarrow \mathbf{0}$
for $t = 1$ to T **do**
  Play $\mathbf{w}_t$ and receive ℓ_t which is L_t-Lipschitz and (σ_t, q_t)-uniformly convex
  Pick λ_t that satisfies (21)
  $\Phi_{t+1} \leftarrow \Phi_t + \ell_t + \lambda_t \Psi$
  $\mathbf{w}'_{t+1} \leftarrow \nabla \Phi^\star_{t+1}(\nabla \Phi_t(\mathbf{w}_t))$
  Update $\mathbf{w}_{t+1} \leftarrow \underset{\mathbf{w} \in \mathcal{W}}{\operatorname{argmin}}\ \Delta_{\Phi_{t+1}}\left(\mathbf{w}, \mathbf{w}'_{t+1}\right)$
end for

Now for the sake of readability, we assume that the supremum in (3) is achieved. The following theorem states that the same adaptive algorithm achieves the upper bound suggested in Theorem 11 for various q-uniformly convex games. Further the algorithm adjusts itself to the scenario when $\mathcal{A}$ plays a different (σ_t, q_t)-uniformly convex function at each round t. To see this let, $\sigma^\star_j = \sigma_j \, \mathbb{1}_{\{q_j < q^* + 1\}}$. In the above algorithm we set at each round λ_t that satisfies,

$$2C\lambda_t = \left(\sum_{i \le t} \frac{\frac{\sigma^\star_i}{M_t^{q_i}}}{\left(\sum_{j \le t} \left[\frac{\sigma^\star_j}{M_t^{q_j}} + \frac{\lambda_j}{M_t^{q^\star}} \right] \right)^{p_i}} \right) + \frac{1}{\left(\sum_{j \le t} \left[\frac{\sigma^\star_j}{M_t^{q_j}} + \frac{\lambda_j}{M_t^{q^\star}} \right] \right)^{p^* - 1}} , \tag{21}$$

where $M_t = L_t + C$.

We have the following theorem which upper bounds the regret of the Adaptive Mirror Descent algorithm.

Theorem 12. *Let $\mathcal{W} = U(\mathfrak{B})$. Let AMD denote the $\mathcal{P}$-strategy obtained by running Adaptive Mirror Descent with a Ψ which is $q^\star = \frac{p^\star(\mathfrak{B}^\star)}{p^\star(\mathfrak{B}^\star) - 1}$ uniformly convex. Then, for all sequences $\ell_{1:T}$ such that ℓ_t is L_t-Lipschitz and (σ_t, q_t)-uniformly convex, we have,*

$$\mathrm{Reg}(AMD, \ell_{1:T}) \le \min_{\lambda_{1:T}} \sum_{t=1}^{T} \left\{ \left(\sum_{i \le t} \frac{\frac{2\sigma^\star_i}{M_t^{q_i}}}{\left(\sum_{j \le t} \left[\frac{\sigma^\star_j}{M_t^{q_j}} + \frac{\lambda_j}{M_t^{q^\star}} \right] \right)^{p_i}} \right) + \frac{2}{\left(\sum_{j \le t} \left[\frac{\sigma^\star_j}{M_t^{q_j}} + \frac{\lambda_j}{M_t^{q^\star}} \right] \right)^{p^* - 1}} + 2\lambda_t C \right\}$$

Proof. First note that $f_t = \ell_t + \lambda_t \Psi$ and further Ψ is a $q^\star$-uniformly convex and ℓ_t is (σ_t, q_t)-uniformly convex. Hence we see that

$$\begin{aligned} \Delta_{f_t}\left(\mathbf{w}'_{t+1}, \mathbf{w}_t\right) &\ge \frac{\sigma_t}{q_t} \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^{q_t} + \frac{\lambda_t}{q^\star} \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^{q^\star} \\ &\ge \frac{\sigma^*_t}{q_t} \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^{q_t} + \frac{\lambda_t}{q^\star} \|\mathbf{w}'_{t+1} - \mathbf{w}_t\|^{q^\star} \end{aligned}$$

Where

$$\sigma^*_t = \begin{cases} \sigma_t & \text{if } q_t \le q^\star + 1 \\ 0 & \text{otherwise} \end{cases}$$

Now since $\Phi_{t+1} = \sum_{i \le t} f_t$, we see that

$$\begin{aligned} \Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) &= \left\langle \nabla \Phi_{t+1}(\mathbf{w}_t) - \nabla \Phi_{t+1}(\mathbf{w}'_{t+1}), \mathbf{w}_t - \mathbf{w}'_{t+1} \right\rangle - \Delta_{\Phi_{t+1}}\left(\mathbf{w}'_{t+1}, \mathbf{w}_t\right) \\ &= \left\langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}'_{t+1} \right\rangle - \Delta_{\Phi_{t+1}}\left(\mathbf{w}'_{t+1}, \mathbf{w}_t\right) \\ &\le \left\langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}'_{t+1} \right\rangle - \sum_{i=1}^{t} \frac{\sigma^*_i}{q_i} \|\mathbf{w}'_{i+1} - \mathbf{w}_i\|^{q_i} - \sum_{i=1}^{t} \frac{\lambda_i}{q^\star} \|\mathbf{w}'_{i+1} - \mathbf{w}_i\|^{q^\star} \end{aligned}$$

Now consider any arbitrary sequence $\beta_1, ..., \beta_{2t}$ of non-negative numbers such that $\sum_{i=1}^{2t} \beta_i = 1$. In this case note that by Fenchel-Young inequality,

$$
\begin{aligned}
\Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) &\leq \sum_{i=1}^{2t} \left\langle \beta_i \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}'_{t+1} \right\rangle - \sum_{i=1}^{t} \frac{\sigma_i^*}{q_i} \|\mathbf{w}'_{i+1} - \mathbf{w}_i\|^{q_i} - \sum_{i=1}^{t} \frac{\lambda_i}{q^\star} \|\mathbf{w}'_{i+1} - \mathbf{w}_i\|^{q^\star} \\
&\leq \sum_{i=1}^{t} \left(\frac{\beta_i^{p_i} \|\nabla f_t(\mathbf{w}_t)\|_\star^{p_i}}{p_i (\sigma_i^*)^{p_i/q_i}} + \frac{\beta_{i+t}^{p^\star} \|\nabla f_t(\mathbf{w}_t)\|_\star^{p^\star}}{p^\star \lambda_i^{p^\star/q^\star}} \right) \\
&\leq \sum_{i=1}^{t} \left(\frac{\beta_i^{p_i} (L_t + \lambda_t C)^{p_i}}{p_i (\sigma_i^*)^{p_i/q_i}} + \frac{\beta_{i+t}^{p^\star} (L_t + \lambda_t C)^{p^\star}}{p^\star \lambda_i^{p^\star/q^\star}} \right) \\
&\leq \sum_{i=1}^{t} \left(\frac{\beta_i^{p_i} (L_t + C)^{p_i}}{p_i (\sigma_i^*)^{p_i/q_i}} + \frac{\beta_{i+t}^{p^\star} (L_t + C)^{p^\star}}{p^\star \lambda_i^{p^\star/q^\star}} \right)
\end{aligned}
$$

In the above we used the fact that since ℓ_t is L_t-Lipschitz and Ψ is C-Lipschitz, $\|\nabla f_t(\mathbf{w}_t)\|_\star \leq (L_t + C)$. We also further assume that $\lambda_t \leq 1$ for any λ. Now choosing

$$
\forall i \leq t,\ \beta_i \propto \frac{\sigma_i^*}{(L_t + C)^{q_i}} \quad \text{and } \forall t < i \leq 2t,\ \beta_i \propto \frac{\lambda_i}{(L_t + C)^{q^\star}}
$$

we see that

$$
\begin{aligned}
\Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) &\leq \sum_{i=1}^{t} \left(\frac{\frac{\sigma_i^*}{(L_t+C)^{q_i}}}{p_i \left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p_i}} + \frac{\frac{\lambda_i}{(L_t+C)^{q^\star}}}{p^\star \left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p^\star}} \right) \\
&\leq \sum_{i=1}^{t} \left(\frac{\frac{\sigma_i^*}{(L_t+C)^{q_i}}}{\left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p_i}} + \frac{\frac{\lambda_i}{(L_t+C)^{q^\star}}}{\left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p^\star}} \right) \\
&\leq \sum_{i=1}^{t} \left(\frac{\frac{\sigma_i^*}{(L_t+C)^{q_i}}}{\left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p_i}} + \frac{\frac{\lambda_i}{(L_t+C)^{q^\star}} + \frac{\sigma_i^*}{(L_t+C)^{q_i}}}{\left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p^\star}} \right) \\
&= \left(\sum_{i=1}^{t} \frac{\frac{\sigma_i^*}{(L_t+C)^{q_i}}}{\left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p_i}} \right) + \frac{1}{\left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p^\star - 1}}
\end{aligned}
$$

where the first step is because $p^\star, p_i \geq 1$. Thus using Lemma 16 we conclude that

$$
\begin{aligned}
\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \leq \sum_{t=1}^{T} \Bigg(&\left(\sum_{i=1}^{t} \frac{\frac{\sigma_i^*}{(L_t+C)^{q_i}}}{\left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p_i}} \right) \\
&+ \frac{1}{\left(\sum_{j=1}^{t} \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}} \right) \right)^{p^\star - 1}} + 2C\lambda_t \Bigg)
\end{aligned}
$$

Now since we choose λ_t's that satisfy Equation 21, using Lemma 17 we conclude that

$$\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \leq 2 \operatorname*{argmin}_{\lambda_1,\ldots,\lambda_T \geq 0} \left\{ \sum_{t=1}^{T} \left(\left(\sum_{i=1}^{t} \frac{\frac{\sigma_i^*}{(L_t+C)^{q_i}}}{\left(\sum_{j=1}^{t}\left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}}\right)\right)^{p_i}} \right) + \frac{1}{\left(\sum_{j=1}^{t}\left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}}\right)\right)^{p^\star - 1}} + 2C\lambda_t \right) \right\}$$

□

Using the above regret bound we get the following corollary showing that the Adaptive Mirror Descent algorithm can be used to achieve all upper bounds on the regret presented in the paper.

Corollary 13. *There exists a Ψ which is $q^\star$-uniformly convex function, and using this function with the Adaptive Mirror Descent (AMD) algorithm, we have the following.*
1. *Regret of AMD for convex-Lipschitz game matches upper bound in* (10).
2 . *Regret of AMD for q-uniformly convex game matches upper bound in* (16).
3 . *For the bounded convex game, there exists a $C > 0$ such that using AMD on $1 - CT^{-\frac{1}{2q^\star}}$ ball achieves the upper bound in* (12) *for the game played on the unit ball.*

Proof. Claim 2 is shown in the constructive proof of the upper bound of Theorem 11, specifically Equation 20. As for claim 1, note that this is the case of linear functions and so it is the same as adversary picking each $\sigma_t = 0$ (and pick any q_t as it is immaterial then). Regret in this case again can be found in the proof of the upper bound of Theorem 11 (Equation 20). and so claim 1 also holds. As for the last claim, given claim 1, it is evident from proof of Theorem 4. □

We also note that when $q^\star = 2$ then for any sequence of $\sigma_1, .., \sigma_T$, AMD enjoys the same guarantee as Algorithm 4 in [11] (see Theorem 4.2 for regret gaurantee).

8 Discussion

In this paper, we considered general Banach spaces. Let us consider a particular case: finite d-dimensional spaces ℓ_d^q for $q > 2$. Since the corresponding q-norms are equivalent to the 2-norm up to a factor of $\sqrt{d}$, we can get a regret bound of $O(\sqrt{dT})$ for the convex-Lipschitz game in these spaces. However, using our results, we can also get $O(T^{1/p})$ rates since the dual space ℓ_d^p with $p = q/(q-1)$ has type p with a dimension independent constant. So, we have the interesting consequence that we can trade-off dimensionality dependence with a worse rate in T.

In future work, we also plan to convert the player strategies given here into implementable algorithms. Online learning algorithms can be implemented in infinite dimensional reproducing kernel Hilbert spaces [19] by exploiting the representer theorem and duality. We can, therefore, hope to implement online learning algorithms in infinite dimensional Banach spaces where some analogue of the representer theorem is available. Der and Lee [20] have made progress in this direction using the notion of *semi-inner products*. For $L_q(\Omega, \mu)$ spaces with q even, they showed how the problem of finding a maximum margin linear classifier can be reduced to a finite dimensional convex program using "moment functions". The types (and their associated constants) of L_q spaces are well known from classical Banach space theory. So, we can use their ideas to get online algorithms in these spaces with provable regret guarantees. Vovk [10] also defines "Banach kernels" for certain Banach spaces of real valued functions and gives an implementable algorithm assuming the Banach kernel is efficiently computable. His interest is in prediction with the squared loss. It will be interesting to explore the connection of his ideas with the setting of this paper.

Using online-to-batch conversions, our results also imply error bounds for the estimation error in the batch setting. If $p^\star(\mathfrak{B}^\star) < 2$ then we get a rate worse than $O(T^{-1/2})$. However, we get the ability to work with richer function classes. This can decrease the approximation error. The study of this trade-off can be helpful.

We would also like to improve our lower and/or upper bounds where they do not match. In this regard, we should mention that the upper bound for the convex-bounded game given in Theorem 4 is not tight for a Hilbert space. Our upper bound is $O(T^{3/4})$ but it can be shown that using the self-concordant barrier $\log(1 - \|\mathbf{w}\|^2)$ for the unit ball, we get an upper bound of $O(T^{2/3})$.

References

[1] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[2] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2–3):169–192, 2007.

[3] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications.* PhD thesis, The Hebrew University, 2007.

[4] J. Abernethy, P. L. Bartlett, A. Rakhlin, and A. Tewari. Optimal strategies and minimax lower bounds for online convex games. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2008.

[5] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization.* Nauka Publishers, Moscow, 1978.

[6] J. Abernethy, A. Agarwal, P. L. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the Twentieth Annual Conference on Computational Learning Theory*, 2009.

[7] M. J. Donahue, L. Gurvits, C. Darken, and E. Sontag. Rates of convex approximation in non-Hilbert spaces. *Constructive Approximation*, 13:187–220, 1997.

[8] L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. *Theoretical Computer Science*, 261(1):81–90, 2001.

[9] S. Mendelson and G. Schechtman. The shattering dimension of sets of linear functionals. *The Annals of Probability*, 32(3):1746–1770, 2004.

[10] V. Vovk. Competing with wild prediction rules. *Machine Learning*, 69(2–3):193–212, 2007.

[11] P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 21*, 2007.

[12] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

[13] A. Beck. A convexity condition in Banach spaces and the strong law of large numbers. *Proceedings of the American Mathematical Society*, 13(2):329–334, 1962.

[14] G. Pisier. Sur les espaces de Banach qui ne contiennent pas uniformément de ℓ_n^1. *C. R. Acad. Sci. Paris Sér. A-B*, 277:A991–A994, 1973.

[15] G. Pisier. Probabilistic methods in the geometry of Banach spaces. In G. Letta and M. Pratelli, editors, *Probability and Analysis*, volume 1206 of *Lecture Notes in Mathematics*, pages 167–241. Springer, 1986.

[16] G. Pisier. Martingales with values in uniformly convex spaces. *Israel Journal of Mathematics*, 20(3–4):326–350, 1975.

[17] Y. Gordon and D. R. Lewis. Absolutely summing operators and local unconditional structures. *Acta Mathematica*, 133:27–48, 1974.

[18] W. B. Johnson and J. Lindenstrauss. Basic concepts in the geometry of Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of the Geometry of Banach Spaces*, volume 1, pages 1–84. Elsevier, 2001.

[19] J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8):2165–2176, 2002.

[20] R. Der and D. Lee. Large-margin classification in Banach spaces. In *Processing of Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

[21] C. Zălinescu. *Convex analysis in general vector spaces*. World Scientific Publishing Co. Inc., River Edge, NJ, 2002.

[22] J. Borwein, A. J. Guirao, P. Hájek, and J. Vanderwerff. Uniformly convex functions on Banach spaces. *Proceedings of the American Mathematical Society*, 137(3):1081–1091, 2009.

Appendix

The proof of Theorem 2 required a couple of lemmas which we prove now. The first lemma uses minimax-maximin equality and the idea of *tangent sequences* to prove a general upper bound on the minimax value of a general (not necessarily linear) game. The second lemma gives a simple lower bound argument for linear games.

Lemma 14. *Fix a Banach space $\mathfrak{B}$ and $T \geq 1$. Let $\epsilon_1, \ldots, \epsilon_T$ be i.i.d. Rademacher random variables. If $\mathcal{F}$ is closed under negation then the minimax value satisfies the upper bound,*

$$V\left(\mathcal{W}, \mathcal{F}\right) \leq 2 \cdot \sup_{\ell_{1:T} \in \mathcal{F}^T} \mathbb{E}\left[\sup_{\mathbf{w} \in \mathcal{W}} \left(\sum_{t=1}^{T} \epsilon_t \ell_t(\mathbf{w})\right)\right] .$$

Proof. Recall that we denote a general distribution over $\mathcal{A}$'s sequences by Q and $\mathcal{P}$-strategies by W. Equation (1) gives us,

$$V\left(\mathcal{W}, \mathcal{F}\right) = \sup_{Q} \inf_{W} \mathbb{E}_{\ell_{1:T} \sim Q}\left[\mathrm{Reg}(W, \ell_{1:T})\right] .$$

If we define

$$V_Q := \inf_{W} \mathbb{E}_{\ell_{1:T} \sim Q}\left[\mathrm{Reg}(W, \ell_{1:T})\right] ,$$

this can be written succinctly as,

$$V\left(\mathcal{W}, \mathcal{F}\right) = \sup_{Q} V_Q . \tag{22}$$

Now, let us fix a distribution Q and denote the conditional expectation w.r.t. $\ell_{1:t}$ by $\mathbb{E}_t\left[\cdot\right]$ and the full expectation w.r.t. $\ell_{1:T}$ by $\mathbb{E}\left[\cdot\right]$. Substituting the definition of regret and noting that the infimum in its definition does not depend on the strategy W, we get

$$V_Q = \inf_{W}\left(\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(W_t(\ell_{1:t-1}))\right]\right) - \mathbb{E}\left[\inf_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} \ell_t(\mathbf{w})\right] . \tag{23}$$

Let us simplify the infimum over $\mathcal{P}$-strategies as follows,

$$\begin{aligned}
\inf_{W} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(W_t(\ell_{1:t-1}))\right] &= \inf_{W} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_{t-1}\left[\ell_t(W_t(\ell_{1:t-1}))\right]\right] \\
&= \sum_{t=1}^{T} \inf_{W_t} \mathbb{E}\left[\mathbb{E}_{t-1}\left[\ell_t(W_t(\ell_{1:t-1}))\right]\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\inf_{\mathbf{w}_t \in \mathcal{W}} \mathbb{E}_{t-1}\left[\ell_t(\mathbf{w}_t)\right]\right] .
\end{aligned}$$

Substituting this into (23), we get,

$$\begin{aligned}
V_Q &= \sum_{t=1}^{T} \mathbb{E}\left[\inf_{\mathbf{w}_t \in \mathcal{W}} \mathbb{E}_{t-1}\left[\ell_t(\mathbf{w}_t)\right]\right] - \mathbb{E}\left[\inf_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} \ell_t(\mathbf{w})\right] \\
&= \mathbb{E}\left[\sup_{\mathbf{w} \in \mathcal{W}} \sum_{t=1}^{T} -\ell_t(\mathbf{w})\right] - \sum_{t=1}^{T} \mathbb{E}\left[\sup_{\mathbf{w}_t \in \mathcal{W}} \mathbb{E}_{t-1}\left[-\ell_t(\mathbf{w}_t)\right]\right] . \qquad (24)
\end{aligned}$$

Up to this point we have equality. Now, we use subadditivity of "sup" to get,

$$
\begin{aligned}
V_Q &\leq \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \left(-\ell_t(\mathbf{w}) - \mathbb{E}_{t-1}\left[-\ell_t(\mathbf{w})\right]\right)\right] \\
&\quad + \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \mathbb{E}_{t-1}\left[-\ell_t(\mathbf{w})\right]\right] - \sum_{t=1}^{T} \mathbb{E}\left[\sup_{\mathbf{w}_t\in\mathcal{W}} \mathbb{E}_{t-1}\left[-\ell_t(\mathbf{w}_t)\right]\right] \qquad (25) \\
&\leq \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \left(-\ell_t(\mathbf{w}) - \mathbb{E}_{t-1}\left[-\ell_t(\mathbf{w})\right]\right)\right] . \qquad (26)
\end{aligned}
$$

Note that $-\ell_t(\mathbf{w}) - \mathbb{E}_{t-1}\left[-\ell(\mathbf{w})\right]$ is a martingale difference sequence. Now, we want to use an analogue of the *symmetrization* technique used in empirical process theory. There, symmetrization is achieved by introducing a "ghost sample", i.e. another i.i.d. sequence independent of the first one and having the same distribution as the first. Here, instead, we will need a *tangent sequence* $\tilde{\ell}_{1:T}$ that is constructed as follows. Let $\tilde{\ell}_1$ be an independent copy of ℓ_1. For $t \geq 2$, let $\tilde{\ell}_t$ have the same conditional distribution as ℓ_t given $\ell_{1:t-1}$ also be independent of ℓ_t given $\ell_{1:t-1}$. Let $\mathcal{G}$ denote the sigma field generated by $\ell_{1:T}$. Then, we have,

$$
\mathbb{E}_{t-1}\left[\ell_t\right] = \mathbb{E}_{t-1}\left[\tilde{\ell}_t\right] = \mathbb{E}_{\mathcal{G}}\left[\tilde{\ell}_t\right] . \qquad (27)
$$

The first equality is true by construction. The second holds because $\tilde{\ell}_t$ is conditionally independent of $\ell_{t:T}$ given $\ell_{1:t-1}$. We also have,

$$
\ell_t = \mathbb{E}_{\mathcal{G}}\left[\ell_t\right] \qquad (28)
$$

because ℓ_t is $\mathcal{G}$-measurable. Plugging in (27) and (28) into (26), we get,

$$
\begin{aligned}
V_Q &\leq \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \left(\mathbb{E}_{\mathcal{G}}\left[-\ell_t(\mathbf{w})\right] - \mathbb{E}_{\mathcal{G}}\left[-\tilde{\ell}_t(\mathbf{w})\right]\right)\right] \\
&= \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \mathbb{E}_{\mathcal{G}}\left[\sum_{t=1}^{T} \tilde{\ell}_t(\mathbf{w}) - \ell_t(\mathbf{w})\right]\right] \\
&\leq \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \tilde{\ell}_t(\mathbf{w}) - \ell_t(\mathbf{w})\right] , \qquad (29)
\end{aligned}
$$

where the last inequality follows by Jensen's. At this stage, the expectation is over $\ell_{1:T}$ and $\tilde{\ell}_{1:T}$.

Now introduce i.i.d. Rademacher random variables $\epsilon_1, \ldots, \epsilon_T$ that are also independent of all the random variables defined so far. For any $t > 1$, conditioned on $\ell_{1:t-1}$, we have (where "$\stackrel{d}{=}$" denotes equality in distribution),

- $(\ell_t, \tilde{\ell}_t) \stackrel{d}{=} (\tilde{\ell}_t, \ell_t)$, and
- $(\ell_t, \tilde{\ell}_t)$ is independent of $\tilde{\ell}_{1:t-1}$.

Thus, conditioned on $\ell_{1:t-1}$,

$$
(\tilde{\ell}_{1:t-1}, \ell_t, \tilde{\ell}_t) \stackrel{d}{=} (\tilde{\ell}_{1:t-1}, \tilde{\ell}_t, \ell_t) .
$$

Therefore,

$$
(\ell_{1:t-1}, \tilde{\ell}_{1:t-1}, \ell_t, \tilde{\ell}_t) \stackrel{d}{=} (\ell_{1:t-1}, \tilde{\ell}_{1:t-1}, \tilde{\ell}_t, \ell_t) . \qquad (30)
$$

This is a crucial distributional equality and we will use it repeatedly to introduce the Rademacher variables into (29) one by one. Using (30) with $t = T$, we can rewrite (29) as,

$$
V_Q \leq \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T-1} (\tilde{\ell}_t(\mathbf{w}) - \ell_t(\mathbf{w})) + \epsilon_T(\tilde{\ell}_T(\mathbf{w}) - \ell_T(\mathbf{w}))\right] .
$$

Here, the expectation is over $\ell_{1:T}$, $\tilde{\ell}_{1:T}$ and ϵ_T. We now get rid of the expectation over $\ell_T, \tilde{\ell}_T$ by upper bounding it by the supremum over ℓ'_T, ℓ''_T in $\mathcal{F}$. Thus,

$$V_Q \leq \sup_{\ell'_T, \ell''_T \in \mathcal{F}} \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T-1} (\tilde{\ell}_t(\mathbf{w}) - \ell_t(\mathbf{w})) + \epsilon_T(\ell''_T(\mathbf{w}) - \ell'_T(\mathbf{w}))\right] .$$

In the above expression, the expectation is over $\ell_{1:T-1}$, $\tilde{\ell}_{1:T-1}$ and ϵ_T.

Now repeating this argument by using (30) for $t = T-1, \ldots, 1$ (in that order), we finally get,

$$V_Q \leq \sup_{\ell'_{1:T}, \ell''_{1:T} \in \mathcal{F}^T} \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \epsilon_t(\ell''_t(\mathbf{w}) - \ell'_t(\mathbf{w}))\right] .$$

At this stage, the expectation is only over the Rademacher random variables. Finally, using subdditivity of "sup" and the fact that $\mathcal{F}$ is closed under negation, we get

$$\begin{aligned} V_Q &\leq \sup_{\ell''_{1:T} \in \mathcal{F}^T} \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \epsilon_t \ell''_t(\mathbf{w})\right] + \sup_{\ell'_{1:T} \in \mathcal{F}^T} \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \epsilon_t \ell'_t(\mathbf{w})\right] \\ &= 2 \sup_{\ell'_{1:T} \in \mathcal{F}^T} \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \epsilon_t \ell_t(\mathbf{w})\right] . \end{aligned}$$

Plugging this into (22) finishes the proof. □

Lemma 15. *Fix a Banach space $\mathfrak{B}$ and $T \geq 1$. If $\mathcal{F}$ consists only of linear functions on $\mathcal{W}$ and is closed under negation, then*

$$\sup_{\ell_{1:T} \in \mathcal{F}^T} \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \left(\sum_{t=1}^{T} \epsilon_t \ell_t(\mathbf{w})\right)\right] \leq V(\mathcal{W}, \mathcal{F}) .$$

Proof. Conider the following distribution Q for $\mathcal{A}$. Fix $\ell'_1, \ldots, \ell'_T \in \mathcal{F}$ and choose i.i.d. Rademacher random variables $\epsilon_1, \ldots, \epsilon_T$. Set $\ell_t = \epsilon_t \ell'_t$. Under this distribution, for any $\mathbf{w}_t \in \mathcal{W}$, $\mathbb{E}_{t-1}[-\ell_t(\mathbf{w}_t)] = -\ell'_t(\mathbf{w}_t)\mathbb{E}_{t-1}[\epsilon_t] = 0$. Thus, by (24), we have,

$$V_Q = \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} -\epsilon_t \ell'_t(\mathbf{w})\right] .$$

Thus, we have,

$$\begin{aligned} V(\mathcal{W}, \mathcal{F}) &= \sup_{Q} V_Q \\ &\geq \sup_{\ell'_{1:T} \in \mathcal{F}^T} \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} -\epsilon_t \ell'_t(\mathbf{w})\right] \\ &= \sup_{\ell'_{1:T} \in \mathcal{F}^T} \mathbb{E}\left[\sup_{\mathbf{w}\in\mathcal{W}} \sum_{t=1}^{T} \epsilon_t \ell'_t(\mathbf{w})\right] , \end{aligned}$$

where the last step used the fact that $\mathcal{F}$ is closed under negation. □

Lemma 16. *For the Adaptive Mirror Descent Algorithm we have that*

$$\mathrm{Reg}(\mathbf{w}_{1:T}, \ell_{1:T}) \leq \sum_{t=1}^{T} \left(\Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) + 2\lambda_t C\right)$$

Proof. Let $f_t = \ell_t + \lambda_t \Psi$, then by definition of Bregman divergence we see that for any $\mathbf{w}^\star \in \mathcal{W}$ we have that,

$$f_t(\mathbf{w}_t) - f_t(\mathbf{w}^\star) = \langle \nabla f_t(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}^\star \rangle - \Delta_{f_t}(\mathbf{w}^\star, \mathbf{w}_t)$$

However note that since $\nabla \Phi^\star_{t+1} = (\nabla \Phi_{t+1})^{-1}$ the update step in the algorithm implies that

$$\nabla \Phi_{t+1}(\mathbf{w}'_{t+1}) = \nabla \Phi_t(\mathbf{w}_t) = \nabla \Phi_{t+1}(\mathbf{w}_t) - \nabla f_t(\mathbf{w}_t)$$

Where the last inequality is because $\Phi_{t+1} = \Phi_t + f_t$ and hence,

$$\begin{aligned} f_t(\mathbf{w}_t) - f_t(\mathbf{w}^\star) &= \left\langle \nabla \Phi_{t+1}(\mathbf{w}_t) - \nabla \Phi_{t+1}(\mathbf{w}'_{t+1}), \mathbf{w}^\star - \mathbf{w}_t \right\rangle - \Delta_{f_t}(\mathbf{w}^\star, \mathbf{w}_t) \\ &= \Delta_{\Phi_{t+1}}(\mathbf{w}^\star, \mathbf{w}_t) - \Delta_{\Phi_{t+1}}\left(\mathbf{w}^\star, \mathbf{w}'_{t+1}\right) + \Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) - \Delta_{f_t}(\mathbf{w}^\star, \mathbf{w}_t) \\ &\le \Delta_{\Phi_{t+1}}(\mathbf{w}^\star, \mathbf{w}_t) - \Delta_{\Phi_{t+1}}(\mathbf{w}^\star, \mathbf{w}_{t+1}) + \Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) - \Delta_{f_t}(\mathbf{w}^\star, \mathbf{w}_t) \end{aligned}$$

where the last step is by Pythagoras theorem for Bregman divergences. Hence adding up we find that

$$\begin{aligned} \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{w}^\star)) &\le \sum_{t=2}^T \left(\Delta_{\Phi_{t+1}}(\mathbf{w}^\star, \mathbf{w}_t) - \Delta_{\Phi_t}(\mathbf{w}^\star, \mathbf{w}_t) - \Delta_{f_t}(\mathbf{w}^\star, \mathbf{w}_t)\right) \\ &\qquad + \Delta_{\Phi_2}(\mathbf{w}^\star, \mathbf{w}_1) - \Delta_{f_1}(\mathbf{w}^\star, \mathbf{w}_1) + \sum_{t=1}^T \Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) \\ &= \sum_{t=1}^T \Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) \end{aligned}$$

The last inequality is due to linearity of Bregman divergence and the fact that $\Phi_{t+1} = \Phi_t + f_t$. Also $\Phi_2 = f_2$. Thus we have that for any $\mathbf{w}^\star \in \mathcal{W}$,

$$\begin{aligned} \sum_{t=1}^T (\ell_t(\mathbf{w}_t) - \ell_t(\mathbf{w}^\star)) &\le \sum_{t=1}^T \left(\Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) + \lambda_t \Psi(\mathbf{w}^\star) - \lambda_t \Psi(\mathbf{w}_t)\right) \\ &\le \sum_{t=1}^T \left(\Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) + \lambda_t C \|\mathbf{w}^\star - \mathbf{w}_t\|\right) \\ &\le \sum_{t=1}^T \left(\Delta_{\Phi_{t+1}}\left(\mathbf{w}_t, \mathbf{w}'_{t+1}\right) + 2\lambda_t C\right) \end{aligned}$$

where the last two inequalities are because $\mathcal{W} = U(\mathfrak{B})$ and the function Ψ is C-Lipschitz on the unit ball. □

Lemma 17. *Define for any sequence $\lambda_1, ..., \lambda_S$ of any size S,*

$$O_S(\lambda_1, ..., \lambda_S) = \sum_{t=1}^S \left(\sum_{i=1}^t \left(\frac{\frac{\sigma_i^*}{(L_t+C)^{q_i}}}{\left(\sum_{j=1}^t \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}}\right)\right)^{p_i}} \right) + \frac{1}{\left(\sum_{j=1}^t \left(\frac{\sigma_j^*}{(L_t+C)^{q_j}} + \frac{\lambda_j}{(L_t+C)^{q^\star}}\right)\right)^{p^\star - 1}} + 2C\lambda_t \right)$$

Then as long as we pick λ_t that satisfies Equation 21, we have that for any T

$$O_T(\lambda_1, ..., \lambda_T) \le 2 \min_{\lambda_1, ..., \lambda_T} O\{\lambda_1, ..., \lambda_T\}$$

Proof. We prove this by induction on S. First consider the case when $S = 1$ and any $\lambda^\star \ge 0$. Now there are two cases, first if $\lambda_1 \le \lambda_1^\star$, then note that

$$O_1(\lambda_1) = 4C\lambda_1 \le 4C\lambda_1^\star \le 2O\{\lambda_1^\star\}$$

On the other hand if $\lambda_1 > \lambda_1^\star$, then

$$O_1(\lambda_1) = \frac{2\frac{\sigma_1^*}{(L_1+C)^{q_1}}}{\left(\frac{\sigma_1^*}{(L_1+C)^{q_1}} + \frac{\lambda_1}{(L_1+C)^{q^\star}}\right)^{p_1}} + \frac{2}{\left(\frac{\sigma_1^*}{(L_1+C)^{q_1}} + \frac{\lambda_1}{(L_1+C)^{q^\star}}\right)^{p^\star-1}}$$

$$\leq \frac{2\frac{\sigma_1^*}{(L_1+C)^{q_1}}}{\left(\frac{\sigma_1^*}{(L_1+C)^{q_1}} + \frac{\lambda_1^\star}{(L_1+C)^{q^\star}}\right)^{p_1}} + \frac{2}{\left(\frac{\sigma_1^*}{(L_1+C)^{q_1}} + \frac{\lambda_1^\star}{(L_t+C)^{q^\star}}\right)^{p^\star-1}} \leq 2O_1(\lambda_1^\star)$$

Thus for $S = 1$ we have the required statement. Now assume the inductive hypothesis hold for $S = T-1$, that is we have that $O_{T-1}(\lambda_1, ..., \lambda_{T-1}) \leq 2\min_{\lambda_1^\star,...,\lambda_{T-1}^\star} O_{T-1}(\lambda_1^\star, ..., \lambda_{T=1}^\star)$. Now consider any $\lambda_1^\star, ..., \lambda_T^\star \geq 0$. We have two possibilities, first $\sum_{t=1}^T \lambda_t \leq \sum_{t=1}^T \lambda_t^\star$. In this case,

$$O_T(\lambda_1, ..., \lambda_T) = 4C\sum_{t=1}^T \lambda_t \leq 4C\sum_{t=1}^T \lambda_t^\star \leq 2O_T(\lambda_1^\star, ..., \lambda_T^\star)$$

Next case is when $\sum_{t=1}^T \lambda_t > \sum_{t=1}^T \lambda_t^\star$. In this case from Equation 21 see that

$$\begin{aligned}
&2C\lambda_T + \sum_{i=1}^T \left(\frac{\frac{\sigma_i^*}{(L_T+C)^{q_i}}}{\left(\sum_{j=1}^T \left(\frac{\sigma_j^*}{(L_T+C)^{q_j}} + \frac{\lambda_j}{(L_T+C)^{q^\star}}\right)\right)^{p_i}} \right) + \frac{1}{\left(\sum_{j=1}^T \left(\frac{\sigma_j^*}{(L_T+C)^{q_j}} + \frac{\lambda_j}{(L_T+C)^{q^\star}}\right)\right)^{p^\star-1}} \\
&= 2\sum_{i=1}^T \left(\frac{\frac{\sigma_i^*}{(L_T+C)^{q_i}}}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \frac{\sum_{j=1}^T \lambda_j}{(L_T+C)^{q^\star}}\right)^{p_i}} \right) + \frac{2}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \frac{\sum_{j=1}^T \lambda_j}{(L_T+C)^{q^\star}}\right)^{p^\star-1}} \\
&< 2\sum_{i=1}^T \left(\frac{\frac{\sigma_i^*}{(L_T+C)^{q_i}}}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \frac{\sum_{j=1}^T \lambda_j^\star}{(L_T+C)^{q^\star}}\right)^{p_i}} \right) + \frac{2}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \frac{\sum_{j=1}^T \lambda_j^\star}{(L_T+C)^{q^\star}}\right)^{p^\star-1}} \\
&\leq 2\sum_{i=1}^T \left(\frac{\frac{\sigma_i^*}{(L_T+C)^{q_i}}}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \sum_{j=1}^T \frac{\lambda_j^\star}{(L_T+C)^{q^\star}}\right)^{p_i}} \right) + \frac{2}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \sum_{j=1}^T \frac{\lambda_j^\star}{(L_T+C)^{q^\star}}\right)^{p^\star-1}} + 4C\lambda_T^\star
\end{aligned} \tag{31}$$

However note that for any sequence $\lambda_1', ..., \lambda_T'$, by definition

$$\begin{aligned}
O_T(\lambda_1', ..., \lambda_T) =& O_{T-1}(\lambda_1', ..., \lambda_{T-1}') + 2C\lambda_T' \\
&+ \sum_{i=1}^T \left(\frac{\frac{\sigma_i^*}{(L_T+C)^{q_i}}}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \sum_{j=1}^T \frac{\lambda_j'}{(L_T+C)^{q^\star}}\right)^{p_i}} \right) + \frac{1}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \sum_{j=1}^T \frac{\lambda_j'}{(L_T+C)^{q^\star}}\right)^{p^\star-1}}
\end{aligned} \tag{32}$$

Hence applying the above on $\lambda_1, ..., \lambda_T$ used by the algorithm and using inductive hypothesis and Equation 31 we see that

$$\begin{aligned}
O_T(\lambda_1, ..., \lambda_T) \leq& 2\min_{\lambda_1^\star,...,\lambda_{T-1}^\star} O_{T-1}(\lambda_1^\star, ..., \lambda_{T-1}^\star) \\
&+ 2\left(\sum_{i=1}^T \left(\frac{\frac{\sigma_i^*}{(L_T+C)^{q_i}}}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \frac{\sum_{j=1}^T \lambda_j^\star}{(L_T+C)^{q^\star}}\right)^{p_i}} \right) + \frac{1}{\left(\sum_{j=1}^T \frac{\sigma_j^*}{(L_T+C)^{q_j}} + \frac{\sum_{j=1}^T \lambda_j^\star}{(L_T+C)^{q^\star}}\right)^{p^\star-1}} + 2C\lambda_T^\star\right)
\end{aligned}$$

Since the second term holds for any $\lambda_1^\star, ..., \lambda_T^\star$ and from Equation 32 we can conclude that

$$O_T(\lambda_1, ..., \lambda_T) \leq 2 \min_{\lambda_1^\star, ..., \lambda_T^\star \geq 0} O_T(\lambda_1^\star, ..., \lambda_T^\star)$$

Thus we see that by induction the statement of the lemma is true. $\square$