Appendix Organization

This appendix is organized as follows:

1. In Section A, we provide our results generalizing the constrained hints algorithm of Section 3 to the general q-uniform convex case, where $q \geq 2$.
2. In Section B, we provide some background on the FTRL framework and extend the literature on adaptive FTRL to Banach spaces.
3. In Section C, we provide a lower bound for the q-uniformly convex case showing that even if a regret bound is allowed to be non-dimension free and *all* of the hints are good, it is not possible to achieve logarithmic regret for general $q > 2$.

A. Constrained Online Learning in q-Uniformly Convex Space

A.1. Preliminaries and notation

We first establish some notation about convex functions and spaces. We say that f is (q, μ)-strongly convex with respect to the norm $\|\cdot\|$ if for all x, y and $g \in \partial f(x)$, we have $f(y) \geq f(x) + \langle g, y-x\rangle + \frac{\mu}{q}\|x-y\|^q$. We say that the Banach space $\mathbb{B}$ is q-uniformly convex if the function $\frac{1}{q}\|x\|^q$ is (q,μ)-strongly convex for some μ. We note this notion is equivalent to the definition of q-uniform convexity of a space used in Dekel et al. (2017) (e.g., see the discussion after Definition 4.16 in Pisier (2016)). Throughout this section, we assume that $\mathbb{B}$ is reflexive and q-uniformly convex with $q \geq 2$. We define p such that $\frac{1}{p} + \frac{1}{q} = 1$.

We also slightly modify the definitions of $G_{T,\alpha}$ and $B_{T,\alpha}$:

$$G_{T,\alpha} = \{t \leq T : \langle c_t, h_t\rangle \geq \alpha \cdot \|c_t\|_*^p\}, \qquad \text{and}$$
$$B_{T,\alpha} = \{t \leq T : \langle c_t, h_t\rangle < \alpha \cdot \|c_t\|_*^p\}$$

A.2. General $q \geq 2$ algorithm and analysis

Our approach for for this general (q,μ)-strongly convex case is essentially the same as in the case when $q = 2$: we use a base algorithm $\mathcal{A}$ to produce points $\bar{x}_t$, and then we augment these points with the hint h_t to play $x_t = \bar{x}_t - \delta_r(x_t)h_t$. However, we require a slightly different definition of δ_r, that generalizes the previous analysis for $q = 2$:

$$\delta_r(x) = \frac{1}{qr}(1 - \|x\|^q).$$

We show that $x - \delta_r(x)h_t \in \mathbb{K}$ for all $x \in \mathbb{K}$, just as we did for the $q = 2$ case in the main text:

Lemma A.1. *For any $r \geq 1$, $\|x\| \leq 1$, and $\|h\| \leq 1$ we have*

$$\|x - \delta_r(x)h_t\| \leq 1.$$

Proof. We proceed by triangle inequality:

$$\begin{aligned}\|x - \delta_r(x)h_t\| &\leq \|x\| + |\delta_r(x)|\|h\| \\ &\leq \|x\| + \frac{1-\|x\|^q}{qr} \\ &\leq \|x\| + \frac{1-\|x\|^q}{q} \\ &\leq \sup_{z\in[0,1]} z + \frac{1-z^q}{q}.\end{aligned}$$

Now observe that the derivative of $z + \frac{1-z^q}{q}$ is $1 - z^{q-1}$, which is positive for all $z \in [0,1]$ and $q \geq 1$. Therefore, the supremum occurs at $z = 1$, for which the value is 1. □

Next, we introduce our expression for the surrogate loss ℓ, which is identical to its previous form:

$$\ell_{h,r,c}(x) = \langle c, x\rangle - |\langle c, h\rangle|\delta_r(x).$$

We can verify the following properties of the surrogate loss, again using essentially the same arguments as for the $q = 2$ case:

Lemma A.2. *Suppose $\mathbb{B}$ is q-uniformly convex for some $q \geq 1$. Let $\|h\| \leq 1$, $\|c\|_* \leq 1$, and $r \geq 1$. If $\langle c, h\rangle \geq 0$, then for all x and u in $\mathbb{K}$, we have*

$$\langle c, x - \delta_r(x)h - u\rangle \leq \ell_{h,r,c}(x) - \ell_{h,r,c}(u).$$

Next, even if $\langle c, h\rangle < 0$, then for all x and u in $\mathbb{K}$, we still have

$$\langle c, x - \delta_r(x)h - u\rangle \leq \ell_{h,r,c}(x) - \ell_{h,r,c}(u) + \frac{2|\langle c, h\rangle|}{qr}.$$

Finally, $\ell_{h,r,c}(x)$ is $\left(q, \frac{|\langle c,h\rangle|\mu}{r}\right)$-strongly convex and $2\|c\|_*$-Lipschitz on $\mathbb{K}$, regardless of the value of $\langle c, h\rangle$.

Proof. First, we notice that since $\|x\| \leq 1$ and $\delta \geq 0$, we must have $\ell_{h,r,c}(u) \leq \langle c, u\rangle$ regardless of the value of $\langle c, h\rangle$. Next, we consider two cases, either $\langle c, h\rangle \geq 0$ or not.

In the former case, $\langle c, h\rangle = |\langle c, h\rangle|$ so that by definition $\ell_{h,r,c}(x) = \langle c, x - \delta_r(x)h\rangle$. Combined with $\ell_{h,r,c}(u) \leq \langle c, u\rangle$, this implies the desired inequality.

In the latter case, $\ell_{h,r,c}(x) = \langle c, x + \delta_r(x)h\rangle = \langle c, x - \delta_r(x)h\rangle + 2\langle c, h\rangle\delta_r(x)$. To conclude, notice that $\delta_r(x) \leq \frac{1}{qr}$ because $\|x\| \leq 1$, so that $-2\langle c, h\rangle\delta_r(x) \leq \frac{2|\langle c,h\rangle|}{qr}$.

Next, we address strong convexity. By definition of $\ell_{h,r,c}$ and δ_r, we have

$$\ell_{h,r,c}(x) = \langle c, x\rangle + \frac{|\langle c, h\rangle|}{qr}(\|x\|^q - 1).$$

Then since $\mathbb{B}$ is q-uniformly convex, $\frac{1}{q}\|x\|^q$ is (q, μ)-strongly convex with respect to $\|\cdot\|$. Since adding a convex function to a strongly convex function maintains the strong convexity, the strong convexity of $\ell_{h,r,c}$ follows.

Finally, for Lipschitzness, notice that the the function $z \mapsto \frac{z^q}{q}$ is 1-Lipschitz on $[-1, 1]$ for all $q \geq 1$. Therefore $\ell_{h,r,c}$ is $\|c\|_* + \frac{|\langle c,h\rangle|}{r}$-Lipschitz. Then since $\|h\| \leq 1$ and $r \geq 1$, $\frac{|\langle c,h\rangle|}{r} \leq \|c\|_*$ and so we are done. □

Algorithm 3 Constrained Imperfect Hints in (q, μ)-Uniformly Convex Space

```
Input: Strong convexity parameters $q, \mu$, norm $\|\cdot\|$, scalar $\eta$
Define $\lambda_0 = \frac{2}{\mu^{1/p}p^{1/p}}$
Define $\bar{x}_1 = 0$
Define $r_1 = 1$
for $t = 1 \dots T$ do
    Get hint $h_t$
    Set $x_t = \bar{x}_t - \delta_{r_t}(\bar{x}_t)$
    Play $x_t$, receive cost $c_t$
    if $\langle c_t, h_t\rangle < 0$ then
        Set $r_{t+1} = \left(r_t^p + |\langle c_t, h_t\rangle|\frac{1}{\eta^p}\right)^{1/p}$
    else
        Set $r_{t+1} = r_t$
    end if
    Define $\ell_t(x) = \ell_{h_t,r_t,c_t}(x)$
    Define $\sigma_t = \frac{|\langle c_t,h_t\rangle|\mu}{r_t}$
    Define $\lambda_t$ as the solution to:
        $\lambda_t = \frac{2^p}{p}\frac{\|c_t\|_*^p}{(\sigma_{1:t} + \mu\lambda_{1:t})^{p/q}}$
    Set $\bar{x}_{t+1} = \text{argmin}_{\|x\|\leq 1} \ell_{1:t}(x) + \frac{\lambda_{0:t}}{q}\|x\|^q$
end for
```

Theorem A.3. *Suppose $\eta \geq 1$. Recall that B_T is set of indices of the "bad hints" such that $\langle c_t, h_t\rangle < 0$. Define*

$$S = \int_1^{1+\sum_{t\in G_{T,\alpha}}\|c_t\|_*^p} z^{-p/q}\, dz.$$

Then Algorithm 3 guarantees:

$$\begin{aligned}\mathcal{R}_{\mathcal{A}}(u, \vec{c}, T) \leq{}& \frac{2}{(\mu p)^{1/p}} + \frac{2^{p+1}}{p(\alpha\mu)^{p/q}}S \\ &+ 2 + \tfrac{8}{p^{1/p}}\left(\sum_{t\in B_{T,\alpha}}\|c_t\|_*^p\right)^{1/p} \\ &+ 2\left(\eta + \frac{2^p S}{p(\eta\alpha\mu)^{p/q}}\right)\left(\sum_{t\in B_T}|\langle c_t, h_t\rangle|\right)^{1/q}.\end{aligned}$$

Before providing the proof of this Theorem, we take a moment to discuss settings for η and more concrete instantiations of the bound. To gain intuition, we will ignore constants and factors of p or q. Thus, the Theorem says:

$$\begin{aligned}\mathcal{R}_{\mathcal{A}}(u, \vec{c}, T) &= O\left(\frac{S}{(\alpha\mu)^{p/q}} + \left(\sum_{t\in B_{T,\alpha}}\|c_t\|_*^p\right)^{1/p}\right. \\ &\qquad\left. + \left(\eta + \frac{S}{(\eta\alpha\mu)^{p/q}}\right)\left(\sum_{t\in B_T}|\langle c_t, h_t\rangle|\right)^{1/q}\right) \\ &\leq O\left(\frac{S}{(\alpha\mu)^{p/q}} + |B_{T,\alpha}|^{1/p}\right. \\ &\qquad\left. + \left(\eta + \frac{S}{(\eta\alpha\mu)^{p/q}}\right)|B_T|^{1/q}\right).\end{aligned}$$

Next, let us bound S. Notice that since $\|c_t\|_* \leq 1$, we have

$$\begin{aligned}S &= \int_1^{1+\sum_{t\in G_{T,\alpha}}\|c_t\|_*^p} z^{-p/q}\, dz \\ &\leq \begin{cases}\log(1 + \sum_{t\in G_{T,\alpha}}\|c_t\|_*^p) & \text{if } q = 2\\ \frac{q-1}{q-2}(1 + \sum_{t\in G_{T,\alpha}}\|c_t\|_*^p)^{\frac{q-2}{q-1}} & \text{if } q > 2\end{cases} \\ &\leq \begin{cases}\log(1 + T) & \text{if } q = 2\\ \frac{q-1}{q-2}(1 + T)^{\frac{q-2}{q-1}} & \text{if } q > 2.\end{cases}\end{aligned}$$

In the special case that $|B_T| = 0$, this recovers the results of (Dekel et al., 2017) in the $q \geq 2$ setting, but allowing for varying hints. In general when $|B_T| \neq 0$, one would like to set $\eta = O(S^{1/p}/(\mu\alpha)^{1/q})$ to obtain:

$$\begin{aligned}\mathcal{R}_{\mathcal{A}}(u, \vec{c}, T) = O&\left(\frac{S}{(\alpha\mu)^{p/q}} + |B_{T,\alpha}|^{1/p}\right. \\ &\left.+ \frac{S^{1/p}}{(\mu\alpha)^{1/q}}|B_T|^{1/q}\right).\end{aligned}$$

Although the final value of S is unknown at the beginning of the game, we can use a doubling-trick based approach to estimate it on-the-fly. Note that this approach however does require fixing a value of α, which is not required by our previous algorithms.

Proof of Theorem A.3. Notice that $r_t \geq 1$ for all t. Thus by Lemma A.2, we have

$$\begin{aligned}\sum_{t=1}^T \langle c_t, \bar{x}_t - \delta_{r_t}(\bar{x}_t)h_t - u\rangle \leq{}& \sum_{t=1}^T \ell_t(\bar{x}_t) - \ell_t(u) \\ &+ \sum_{t\in B_T}\frac{2|\langle c_t, h_t\rangle|}{qr_t}.\end{aligned}$$

First, we will control the last sum in this expression. Observe that by definition, and since $\eta \geq 1$ and $|\langle c_t, h_t\rangle| \leq 1$ for all t, we have

$$r_t = \left(1 + \frac{1}{\eta^p} \sum_{\tau \in B_{t-1}} |\langle c_\tau, h_\tau\rangle|\right)^{1/p}$$
$$\geq \frac{1}{\eta}\left(\sum_{\tau \in B_t} |\langle c_\tau, h_\tau\rangle|\right)^{1/p}.$$

Let $B_T = \{t_1, \dots, t_N\}$. Then using Corollary B.13 we have

$$\sum_{t \in B_T} \frac{|\langle c_t, h_t\rangle|}{r_t} \leq \eta \sum_{i=1}^{N} \frac{|\langle c_{t_i}, h_{t_i}\rangle|}{\left(\sum_{j=i}^{t} |\langle c_{t_j}, h_{t_j}\rangle|\right)^{1/p}}$$
$$\leq \eta q \left(\sum_{t \in B_T} |\langle c_t, h_t\rangle|\right)^{1/q}.$$

So putting this together we have

$$\sum_{t \in B_T} \frac{2|\langle c_t, h_t\rangle|}{q r_t} \leq 2\eta \left(\sum_{t \in B_T} |\langle c_t, h_t\rangle|\right)^{1/q}.$$

Now we turn to bounding $\sum_{t=1}^{T} \ell_t(\bar{x}_t) - \ell_t(u)$. Observe that by Lemma A.2, we have ℓ_t is (q, σ_t)-strongly convex. Therefore, by Theorem B.9, we have

$$\sum_{t=1}^{T} \ell_t(\bar{x}_t) - \ell_t(u) \leq \lambda_{0:T}\|u\|^q + \frac{1}{p}\sum_{t=1}^{T} \frac{\|g_t\|_*^p}{(\sigma_{1:t} + \mu\lambda_{0:t-1})^{p/q}},$$

where $g_t \in \partial \ell_t(\bar{x}_t)$. Then, again by Lemma A.2, ℓ_t is $2\|c_t\|_*$-Lipschitz, so that $\|g_t\|_* \leq 2\|c_t\|_* \leq 2$. Using this fact and $\|u\| \leq 1$, we can write

$$\sum_{t=1}^{T} \ell_t(\bar{x}_t) - \ell_t(u) \leq \lambda_{0:T} + \frac{2^p}{p}\sum_{t=1}^{T} \frac{\|c_t\|_*^p}{(\sigma_{1:t} + \mu\lambda_{0:t-1})^{p/q}}.$$

Next, by Corollary B.11, we have

$$\sum_{t=1}^{T} \ell_t(\bar{x}_t) - \ell_t(u) \leq 2 \inf_{\{\lambda_t^\star\}} \lambda_{1:T}^\star + \frac{2^p}{p}\sum_{t=1}^{T} \frac{\|c_t\|_*^p}{(\sigma_{1:t} + \mu\lambda_{1:t}^\star)^{p/q}}$$
$$+ \frac{2}{(\mu p)^{1/p}}.$$

We upper bound the infimum of λ_t^* by considering only settings where $\lambda_t^* = 0$ for $t > 1$ and $\lambda_1^* \geq \alpha$. Further, we split the sum in the second term into two parts: the indices in $B_{T,\alpha}$ those in $G_{T,\alpha}$. For the indices in $B_{T,\alpha}$, we ignore the influence of the σ_t. For those in $G_{T,\alpha}$, we use the bound $\lambda_1^* \geq \alpha$. This yields:

$$\leq 2 \inf_{\lambda \geq \alpha} \lambda + \frac{2^p}{p\lambda^{p/q}} \sum_{t \in B_{T,\alpha}} \|c_t\|_*^p$$
$$+ \frac{2}{p^{1/p}} + \frac{2^{p+1}}{p} \sum_{t \in G_{T,\alpha}} \frac{\|c_t\|_*^p}{(\mu\alpha + \sigma_{1:t})^{p/q}}.$$

Now by Lemma B.14, we obtain:

$$\inf_{\lambda \geq 1} \lambda + \frac{2^p}{p\lambda^{p/q}} \sum_{t \in B_{T,\alpha}} \|c_t\|_*^p \leq 1 + \tfrac{4}{p^{1/p}} \left(\sum_{t \in B_{T,\alpha}} \|c_t\|_*^p\right)^{1/p}.$$

Next, we observe that since r_t is non-decreasing, we have $\sigma_t \geq \frac{|\langle c_t, h_t\rangle|\mu}{r_T}$. Further, for any $t \in G_{T,\alpha}$, we have by definition $\langle c_t, h_t\rangle \geq \alpha\|c_t\|_*^p$ so that $\sigma_t \geq \frac{\alpha\mu\|c_t\|_*^p}{r_T}$, all of which implies:

$$\sum_{t \in G_{T,\alpha}} \frac{\|c_t\|_*^p}{(\mu\alpha + \sigma_{1:t})^{p/q}}$$
$$\leq \frac{1}{(\alpha\mu)^{p/q}} \sum_{t \in G_{T,\alpha}} \frac{\|c_t\|_*^p r_T^{p/q}}{(1 + \sum_{\tau \in G_{t,\alpha}} \|c_\tau\|_*^p)^{p/q}}.$$

Now invoke Lemma B.12 with $h(z) = z^{-p/q}$ to bound:

$$\sum_{t \in G_{T,\alpha}} \frac{\|c_t\|_*^p}{(1 + \sum_{\tau \notin B_t} \|c_\tau\|_*^p)^{p/q}} \leq \int_1^{1+\sum_{t \in G_{T,\alpha}} \|c_t\|_*^p} z^{-p/q}\, dz$$
$$= S.$$

Next, recall that we have

$$r_T^{p/q} = \left(1 + \frac{1}{\eta^p} \sum_{\tau \in B_{T-1}} |\langle c_\tau, h_\tau\rangle|\right)^{1/q}$$
$$\leq 1 + \frac{1}{\eta^{p/q}} \left(\sum_{\tau \in B_T} |\langle c_\tau, h_\tau\rangle|\right)^{1/q},$$

so that we have

$$\frac{2^{p+1}}{p} \sum_{t \in G_{T,\alpha}} \frac{\|c_t\|_*^p}{(\mu\alpha + \sigma_{1:t})^{p/q}} \leq \frac{2^{p+1}}{p(\alpha\mu)^{p/q}} S r_T^{p/q}$$
$$\leq \frac{2^{p+1}}{p(\alpha\mu)^{p/q}} S + \frac{2^{p+1} S}{\eta^{p/q} p(\alpha\mu)^{p/q}} \left(\sum_{t \in B_T} |\langle c_t, h_t\rangle|\right)^{1/q}.$$

Putting everything we have together so far, we obtain the proof. □

B. Follow-the-Regularized-Leader in Banach Spaces

In this section, we provide some formal definitions and facts in Banach spaces, and generalize prior work on adaptive FTRL algorithms (McMahan, 2017) to the more general q-strongly convex spaces.

Definition B.1. *Given a convex function $f : \mathbb{B} \to \mathbb{R}$, the* Fenchel conjugate *$f^\star : \mathbb{B}^* \to \mathbb{R}$ is defined by $f^*(\theta) = \sup_x \langle \theta, x\rangle - f(x)$.*

Definition B.2. *A Banach space $\mathbb{B}$ is* reflexive *if the map $i : \mathbb{B} \mapsto \mathbb{B}^{**}$ defined by $\langle i(x), \alpha\rangle = \langle \alpha, x\rangle$ is an isomorphism of Banach spaces. When $\mathbb{B}$ is reflexive, we will identify $\mathbb{B}^{**}$ with $\mathbb{B}$ using this isomorphism.*

By the Fenchel–Moreau theorem, $f^* * = f$ whenever $f : \mathbb{B} \to \mathbb{R}$ is convex and lower-semicontinuous and $\mathbb{B}$ is a reflexive Banach space.

Proposition B.3. *Let $\mathbb{B}$ be a reflexive Banach space. Suppose $f : \mathbb{B} \to \mathbb{R}$ is a lower-semicontinuous convex function.*

1. *$f^*(\alpha) = \langle \alpha, a\rangle - f(a)$ if and only if $\alpha \in \partial f(a)$.*
2. *$\alpha \in \partial f(a)$ if and only if $a \in \partial f^*(\alpha)$.*

Proof. 1. Let $h(x)$ be the function defined by $h(x) = f(x) - \langle \alpha, x\rangle$. Notice that $0 \in \partial h(a)$ if and only if a is a minimizer of h, so that $0 \in h(a)$ if and only if $f^*(\alpha) = -h(a)$. Further, $0 \in \partial h(a)$ if and only if $\alpha \in \partial f(a)$. The statement follows.

2. Since $f^{**} = f$, this follows from part 1. □

Definition B.4. *A convex function f is (q, σ)-*strongly convex *with respect to a norm $\|\cdot\|$ if for all x, y and $g \in \partial f(x)$, $f(y) \geq f(x) + \langle g, y - x\rangle + \frac{\sigma}{q}\|x - y\|^q$.*

Definition B.5. *A convex function f is (q, σ)-*strongly smooth *with respect to a norm $\|\cdot\|$ if for all x, y and $g \in \partial f(x)$, $f(y) \leq f(x) + \langle g, y - x\rangle + \frac{\sigma}{q}\|x - y\|^q$.*

Proposition B.6. *Suppose $\mathbb{B}$ is a reflexive Banach space. Let $\frac{1}{p} + \frac{1}{q} = 1$. If $f : \mathbb{B} \to \mathbb{R}$ is (q, σ^q)-strongly convex with respect to a norm $\|\cdot\|$, then $f^* : \mathbb{B} \to \mathbb{R}$ is (p, σ^{-p})-strongly smooth with respect to the dual norm $\|\cdot\|_\star$.*

Proof. Let $\alpha, \beta \in \mathbb{B}^*$ and let $b \in \partial f^*(\beta)$. Define

$$D^* = f^*(\alpha) - f^*(\beta) - \langle \alpha - \beta, b\rangle.$$

It suffices to prove $D^* \leq \frac{1}{p\sigma^p}\|\alpha - \beta\|_\star^p$. By Proposition B.3, we have $\beta \in \partial f(b)$. Let $a \in \partial f^*(\alpha)$ so that $\alpha \in \partial f(a)$. In particular, this implies:

$$f(a) - f(b) - \langle \beta, a - b\rangle \geq \frac{\sigma^q}{q}\|a - b\|^q.$$

We also have:

$$\begin{aligned} f^*(\alpha) &= \langle \alpha, a\rangle - f(a) \\ f^*(\beta) &= \langle \beta, b\rangle - f(b). \end{aligned}$$

Then

$$\begin{aligned} D^* &= \langle \alpha, a\rangle - f(a) - \langle \beta, b\rangle + f(b) - \langle \alpha - \beta, b\rangle \\ &= \langle \alpha, a - b\rangle + f(b) - f(a) \\ &= \langle \alpha - \beta, a - b\rangle + f(b) - f(a) + \langle \beta, a - b\rangle \\ &\leq \langle \alpha - \beta, a - b\rangle - \frac{\sigma^q}{q}\|a - b\|^q \\ &\leq \|\alpha - \beta\|_\star \|a - b\| - \frac{\sigma^q}{q}\|a - b\|^q \\ &\leq \sup_r \|\alpha - \beta\|_\star r - \frac{\sigma^q}{q} r^q \\ &= \frac{1}{p\sigma^p}\|\alpha - \beta\|_\star^p. \end{aligned}$$

□

Next, we prove an analog of McMahan (2017) Lemma 16. The proof is identical, but we use the more general Proposition B.3 and B.6 to verify that it continues to hold in our more general setting.

Lemma B.7. *Suppose $\phi_1 : \mathbb{B} \to \mathbb{R}$ is (q, σ^q) strongly convex with respect to $\mathbb{B}$'s norm $\|\cdot\|$ and let $x_1 =$ argmin ϕ_1. Let $\phi_2(x) = \phi_1(x) + \langle \beta, x\rangle$ for some $\beta \in \mathbb{B}^*$. Then if $x_2 =$ argmin ϕ_2, we have*

$$\phi_2(x_1) - \phi_2(x_2) \leq \frac{1}{p\sigma^p}\|\beta\|_\star^p.$$

Proof. By definition,

$$\begin{aligned} -\phi_1^*(0) &= \inf \phi_1(x) = \phi_1(x_1) \\ -\phi_{.}^*(-\beta) &= -\sup\langle -\beta, x\rangle - \phi_1(x) = \inf \phi_2(x) = \phi_2(x_2). \end{aligned}$$

Now by Proposition B.3 we have $x_1 \in \partial \phi_1^*(0)$ and by Proposition B.6, ϕ_1^* is (p, σ^{-p})-strongly smooth. Therefore:

$$\phi_1^*(-\beta) \leq \phi_1^*(0) - \langle \beta, x_1\rangle + \frac{1}{p\sigma^p}\|\beta\|_\star^p.$$

Then putting all this together we have

$$\begin{aligned} \phi_2(x_1) - \phi_2(x_2) &= \phi_1(x_1) + \langle \beta, x_1\rangle - \phi_2(x_2) \\ &= \phi_1 * (-\beta) - \phi_1^*(0) + \langle \beta, x_1\rangle \\ &\leq \frac{1}{p\sigma^p}\|\beta\|_\star^p. \end{aligned}$$

□

Finally, we have an analog of McMahan (2017) Lemma 7:

Lemma B.8. *Let $\phi_1 : \mathbb{B} \to \mathbb{R}$ be a proper convex function such that $x_1 = \operatorname{argmin} \phi_1(x)$ exists. Let ψ be a convex function such that $\phi_2(x) = \phi_1(x) + \psi_x$ is (q, σ^q)-strongly convex with respect to the norm $\|\cdot\|$. Then for any $\beta \in \partial\psi(x_1)$ and any x_2, we have*

$$\phi_2(x_1) - \phi_2(x_2) \le \frac{1}{p\sigma^p}\|\beta\|_\star^p.$$

Proof. It clearly suffices to prove the result for $x_2 = \operatorname{argmin} \phi_2(x)$. Consider the function $\phi_1'(x) = \phi_2(x) - \langle \beta, x\rangle$. Since $\beta \in \partial\psi(x_1)$, we have $0 \in \partial\phi_1'(x_1)$ so that $x_1 = \operatorname{argmin} \phi_1'(x)$. Further, we clearly have $\phi_2(x) = \phi_1'(x) + \langle \beta, x\rangle$. Therefore by Lemma B.7, we have

$$\phi_2(x_1) - \phi_2(x_2) \le \frac{1}{p\sigma^p}\|\beta\|_\star^p. \qquad \square$$

Now we are ready to state the bound on FTRL, which is an analog of McMahan (2017) Theorem 1 in the more general $q \ge 2$ case.

Theorem B.9. *Suppose $\ell_1, \dots, \ell_T$ are convex functions from $W \to \mathbb{R}$ where $W \subset \mathbb{B}$. Suppose ℓ_t is (q, σ_t)-strongly convex with respect to $\|\cdot\|$. Suppose $\frac{1}{q}\|\cdot\|^q$ is (q, μ)-strongly convex with respect to $\|\cdot\|$. Given arbitrary numbers $\lambda_0, \dots, \lambda_{T-1} > 0$, Define:*

$$\begin{aligned} r_t(x) &= \frac{\lambda_t}{q}\|x\|^q \\ x_{t+1} &= \underset{x}{\mathit{argmin}}\, \ell_{1:t}(x) + r_{0:t}(x). \end{aligned}$$

Let $g_t \in \partial\ell_t(x_t)$. Then we have

$$\begin{aligned} \sum_{t=1}^T \ell_t(x_t) - \ell_t(u) &\le \sum_{t=1}^{T-1} \lambda_t\|u\|^q \\ &\quad + \frac{1}{p}\sum_{t=1}^T \frac{\|g_t\|_\star^p}{(\sigma_{1:t} + \mu\lambda_{0:t-1})^{p/q}}. \end{aligned}$$

Proof. The proof is a nearly immediate consequence of the "Strong FTRL Lemma", (McMahan, 2017) Lemma 5. This result tells us that:

$$\begin{aligned} \sum_{t=1}^T \ell_t(x_t) - \ell_t(u) &\le \sum_{t=1}^{T-1} \lambda_t\|u\|^q \\ &\quad + \sum_{t=1}^T \ell_{1:t}(x_t) + r_{1:t-1}(x_t) - \ell_{1:t}(x_{t+1}) - r_{1:t}(x_{t+1}). \end{aligned}$$

Notice that $x_t = \operatorname{argmin} \ell_{1:t-1}(x) + r_{1:t-1}(x)$. Then observe that $r_t(x_{t+1}) \ge 0$ so that

$$\begin{aligned} &\ell_{1:t}(x_t) + r_{1:t-1}(x_t) - \ell_{1:t}(x_{t+1}) - r_{1:t}(x_{t+1}) \\ &\le \ell_{1:t}(x_t) + r_{1:t-1}(x_t) - \ell_{1:t}(x_{t+1}) - r_{1:t-1}(x_{t+1}) \end{aligned}$$

Finally, we have $\ell_{1:t}(x) + r_{1:t-1}(x)$ is $(q, \sigma_{1:t} + \mu\lambda_{1:t-1})$-strongly convex with respect to $\|\cdot\|$. Therefore applying Lemma B.8 with $\phi_1(x) = \ell_{1:t-1}(x) + r_{1:t-1}(x)$ and $\psi_t(x) = \partial(\ell_t(x_t) + r_{1:t-1}(x_t))$ yields the desired result. $\square$

Next, we need a generalization of Hazan et al. (2008), Lemma 3.1:

Lemma B.10. *Suppose $\lambda_1, \dots, \lambda_T$ is such that*

$$\lambda_t = \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t})^a},$$

for all t for some positive numbers $G_1, \dots, G_T$, $\sigma_1, \dots, \sigma_T$ and a and μ. Then:

$$\sum_{t=1}^T \lambda_t + \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t})^a} \le 2\inf_{\{\lambda_t^\star\}} \sum_{t=1}^T \lambda_t^\star + \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t}^\star)^a}.$$

Proof. The proof is essentially the same as the proof of Lemma 3.1 in Hazan et al. (2008). We proceed by induction. For the base step, consider two cases, either $\lambda_1 \le \lambda_1^\star$ or not. If $\lambda_1 \le \lambda_1^\star$, then we have

$$\lambda_1 + \frac{G_1}{(\sigma_1 + \mu\lambda_1)^a} = 2\lambda_1 \le 2\lambda_1^\star \le 2\lambda_1^\star + 2\frac{G_1}{(\sigma_1 + \mu\lambda_1^\star)^a}.$$

For the other case, when $\lambda_1 > \lambda_1^\star$ we have

$$\begin{aligned} \lambda_1 + \frac{G_1}{(\sigma_1 + \mu\lambda_1)^a} &= 2\frac{G_1}{(\sigma_1 + \mu\lambda_1)^a} \\ &\le 2\frac{G_1}{(\sigma_1 + \mu\lambda_1^\star)^a} \\ &\le 2\lambda_1^\star + 2\frac{G_1}{(\sigma_1 + \mu\lambda_1^\star)^a}. \end{aligned}$$

Now the induction step proceeds in almost exactly the same manner as the base step. Suppose we have

$$\sum_{t=1}^\tau \lambda_t + \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t})^a} \le 2\inf_{\{\lambda_t^\star\}} \sum_{t=1}^\tau \lambda_t^\star + \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t}^\star)^a}.$$

Then consider two cases, either $\lambda_{1:\tau+1} \le \lambda_{1:\tau+1}^\star$ or not. In the first case, we have

$$\begin{aligned} \sum_{t=1}^{\tau+1} \lambda_t + \frac{G_t}{(\sigma_{1:t} + \lambda_{1:t})^a} &= 2\lambda_{1:\tau+1} \\ &\le 2\lambda_{\tau+1}^\star \\ &\le 2\sum_{t=1}^{\tau+1} \lambda_t^\star + \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t}^\star)^a}. \end{aligned}$$

In the other case when $\lambda_{1:\tau+1} > \lambda^\star_{1:\tau+1}$, we have

$$\begin{aligned}\sum_{t=1}^{\tau+1} \lambda_t + \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t})^a} &= 2\sum_{t=1}^{\tau+1} \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t})^a}\\ &\le 2\sum_{t=1}^{\tau+1} \frac{G_t}{(\sigma_{1:t} + \mu\lambda^\star_{1:t})^a}\\ &\le 2\sum_{t=1}^{\tau+1} \lambda^\star_t + \frac{G_t}{(\sigma_{1:t} + \mu\lambda^\star_{1:t})^a}. \square\end{aligned}$$

Finally, a simple corollary of Lemma B.10:

Corollary B.11. *Suppose $\lambda_0, \lambda_1, \dots, \lambda_T$ is such that $\lambda_0 = (M/\mu)^{\frac{1}{a+1}}$ and*

$$\lambda_t = \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t})^a},$$

for $t \ge 1$ for some positive numbers $G_1, \dots, G_T$, $\sigma_1, \dots, \sigma_T$ and a. Then if $G_t \le M$ for all t, we have:

$$\begin{aligned}&\lambda_0 + \sum_{t=1}^{T} \lambda_t + \frac{G_t}{(\sigma_{1:t} + \mu\lambda_{0:t-1})^a}\\ &\le \lambda_0 + 2\inf_{\{\lambda^\star_t\}} \sum_{t=1}^{T} \lambda^\star_t + \frac{G_t}{(\sigma_{1:t} + \mu\lambda^\star_{1:t})^a}.\end{aligned}$$

Proof. The proof is immediate from Lemma B.10, so long as we can establish that $\lambda_t \le \lambda_0$ for all t. To see this, note that $G_t \le M$, so

$$\begin{aligned}\frac{G_t}{(\sigma_{1:t} + \mu\lambda_{1:t})^a} &\le \frac{M}{(\sigma_{1:t} + \mu\lambda_{1:t})^a}\\ &\le \frac{M}{\mu\lambda_t^a}\end{aligned}$$

From this we have $\lambda_t^{a+1} \le \frac{M}{\mu}$, so that $\lambda_t \le (M/\mu)^{\frac{1}{a+1}} = \lambda_0$ as desired. □

We also need the following technical Lemma from Li & Orabona (2019):

Lemma B.12. *Suppose $a_0, \dots, a_T$ are non-negative numbers and $h : [0, \infty) \to [0, \infty)$ is any non-increasing integrable function. Then:*

$$\sum_{t=1}^{T} a_t h(a_{0:t}) \le \int_{a_0}^{a_{0:T}} h(t)\, dt.$$

As special cases of this Lemma, we have:

Corollary B.13. *For any $p > 1$,*

$$\sum_{t=1}^{T} \frac{a_t}{(a_{1:t})^{1/p}} \le q(a_{1:T})^{1/q}.$$

Proof. Set $a_0 = 0$ and $h(z) = \frac{1}{z^{1/p}}$ in Lemma B.12. Hence,

$$\sum_{t=1}^{T} \frac{a_t}{(a_{1:t})^{1/p}} \le \int_0^{a_{1:T}} \frac{dz}{z^{1/p}} = q(a_{1:T})^{1/q}. \qquad \square$$

Finally, we need another technical Lemma:

Lemma B.14. *For all positive real numbers z, A and B and $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\begin{aligned}\inf_{\lambda \ge z} A\lambda + \frac{B}{\lambda^{p/q}} &\le Az + p^{1/p}q^{1/q}A^{1/q}B^{1/p}\\ &\le Az + 2A^{1/q}B^{1/p}.\end{aligned}$$

Proof. Differentiating to solve for the optimal unconstrained λ, we have

$$A - \frac{pB}{q\lambda^{p/q+1}} = 0.$$

Notice that $1 + p/q = p$. Then solving for λ yields:

$$\lambda_\star = \frac{(pB)^{1/p}}{(qA)^{1/p}}.$$

Let us set $\lambda = z + \lambda_\star \ge 1$. Substituting, we have:

$$\begin{aligned}A\lambda + \frac{B}{\lambda^{p/q}} &\le Az + A\lambda_\star + \frac{B}{\lambda_\star^{p/q}}\\ &= Az + \frac{p^{1/p}B^{1/p}A^{1/q}}{q^{1/p}} + \frac{q^{1/q}A^{1/q}B^{1/p}}{p^{1/q}}\\ &= Az + p^{1/p}q^{1/q}A^{1/q}B^{1/p}.\end{aligned}$$

Finally, notice from Young's inequality that $p^{1/p}q^{1/q} \le \frac{p}{p} + \frac{q}{q} = 2$. □

C. Lower bounds for dimension-dependent regret

We now show that a lower bound in the L_q setting even if we allow a dependence on the dimension. Once again, at *every* step, the hints are $\Omega(1)$ correlated with the corresponding cost vectors. In what follows, let $q > 2$ be any real number.

Theorem C.1. *There exists a sequence of hint vectors $h_1, h_2, \dots$ and cost vectors $c_1, c_2, \dots$ in $\mathbb{R}^2$ such that (a) $\langle c_t, h_t \rangle \ge \Omega(1)$ for all t, and (b) any online learning algorithm that plays given the hints incurs an expected regret of $\Omega\left(T^{\frac{(q-2)}{2(q-1)}}\right)$.*

Proof. Again, let e_1, e_2 be an orthonormal basis for $\mathbb{R}^2$, and let $h_t = e_1$ for all t. Let $c_t = e_1 \pm e_2$, where the choice of sign is uniformly random (if c_t are needed to be in the unit ball, we can normalize the vectors; we skip this step as the analysis is identical). Thus for any t, we have $\langle c_t, h_t \rangle = 1 \geq \Omega(1) \left\| c_t \right\|$, for a constant depending only on q (and always $\geq 1/2$).

Now consider any algorithm that plays $\{x_t\}$ within the unit L_q ball. The expected loss is $\sum_t \langle e_1, x_t \rangle$. This is clearly at most T in magnitude. Now, let us consider the best vector in hindsight. Let $z = c_{1:T}$, as before. We have $z = Te_1 + we_2$, for some w of expected magnitude $\sqrt{T}$. We can compute the vector in the L_q ball with the "best" inner product with z. One good choice turns out to be

$$u = \left(1 - \frac{3}{2q} \cdot T^{-\frac{q}{2(q-1)}}\right) e_1 + \text{sign}(w) \cdot T^{-\frac{1}{2(q-1)}} e_2.$$

The fact that $\left\| u \right\|_q \leq 1$ follows using the inequality $(1 - \frac{3\gamma}{2q})^q \leq e^{-3\gamma/2} \leq 1 - \gamma$ for any $\gamma < 1/2$.

For this choice, using the expected magnitude of w,

$$\begin{aligned}\mathbb{E}[\langle z, u \rangle] &= T - \frac{3}{2q} \cdot T^{1 - \frac{q}{2(q-1)}} + T^{\frac{1}{2} - \frac{1}{2(q-1)}} \\ &= T + \left(1 - \frac{3}{2q}\right) T^{\frac{(q-2)}{2(q-1)}}.\end{aligned}$$

Thus for any $q > 2$, the regret is $\Omega(T^{\frac{(q-2)}{2(q-1)}})$, as desired. □