

815 A Experiments

816 In this section, we present a numerical experiment in Figure 1 to validate some of our theoretical
 817 results on the optimality and sub-optimality of fixed and adaptive regularisation for FTRL. We run
 818 FTRL with different regularisers on the loss construction used in the proofs of our lower bounds from
 819 Section 4, which is described in Appendix E.1.

820 For fixed T , we observe that the regret using FTRL with ϕ_p is constant across dimension, while
 821 the regret of FTRL with ϕ_2 increases with dimension. In particular, ϕ_2 outperforms ϕ_p in low-
 822 dimension while ϕ_p outperforms ϕ_2 in high dimensions. This validates our results that ϕ_p is optimal
 823 in high dimensions (Section 2.3) but not in low-dimension (Section 4.2) and that ϕ_2 is optimal in low-
 824 dimension (Section 2.2) but not in high dimensions (Section 4.1). Furthermore, the adaptive procedure
 825 from Section 3 performs well in both low and high dimensions. However, this experiment suggests
 826 that the theoretical threshold $t_0 = 3^{-2p/(p-2)}d$ from Theorem 3.1 is perhaps overly conservative in
 827 the transient setting between low and high dimensions (at least for this loss construction) and that a
 828 larger threshold $t_0 = 2d$ performs better here.

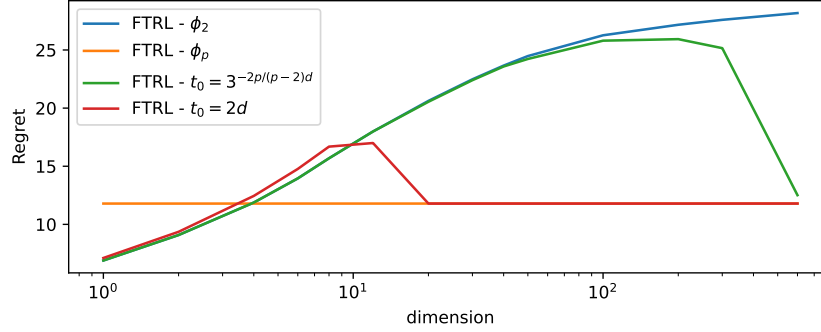


Figure 1: Comparison of FTRL with different regularisation. We fix $T = 40$ (and $L = 1, p = 10$) and vary the dimension. FTRL - ϕ_2 refers to FTRL using the regulariser $\phi_2 = \frac{1}{2}\|x\|_2^2$ from Section 2.2 with $\eta_{t-1} = \sqrt{\frac{d^{1-2/p}}{2t}}$. FTRL - ϕ_p refers to FTRL using the regulariser $\phi_p = \frac{1}{p}\|x\|_p^p$ from Section 2.3 with $\eta_{t-1} = \frac{1}{(2p_*t)^{1/p_*}}$. The final two correspond to using the procedure from Section 3 with adaptive regularisation. The first with the threshold $t_0 = 3^{-2p/(p-2)}d$ from Theorem 3.1, while the second uses the threshold $t_0 = 2d$.

829 **Implementation Details:** The experiment was run on google colab with the default settings
 830 (including CPU) and takes around 5 minutes run. All the details of the loss construction and algorithms
 831 are provided or referenced above. The closed-form updates are provided in Appendix B. Note that
 832 the Bregman projections onto ℓ_p -balls are not available analytically, we use the minimize function
 833 from the scipy.optimize library to compute the projections numerically (with method='SLSQP').

834 **B Closed-form update of FTRL with specific uniformly convex regulariser**
835 **and related lemmas**

836 Consider a regulariser ψ differentiable on \mathbb{R}^d . Define the Bregman divergence of ψ as $D_\psi(x, y) =$
837 $\psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$ for all $x, y \in \mathbb{R}^d$.

838 **Lemma B.1.** Fix $r \geq 2$. Let $\psi(x) = \frac{1}{r} \|x\|_r^r$. Let $V = \mathcal{B}_p$. Let $g_t \in \partial \ell_t(x_t)$. The update rule of
839 FTRL using $\psi_t(x) = \frac{1}{\eta_t - 1} \psi(x)$ as regularisers is

$$\begin{aligned}\tilde{G}_{t+1} &= -\eta_t \sum_{s=1}^t g_s \\ x_{t+1} &= \arg \min_{x \in \mathcal{B}_p} D_\psi \left(x, \text{sign}(\tilde{G}_{t+1}) |\tilde{G}_{t+1}|^{r_\star - 1} \right),\end{aligned}$$

840 where sign , power and absolute value functions are applied component-wise to vectors.

841 *Proof.* Given $g_t \in \partial \ell_t(x_t)$, the update of FTRL with regulariser ψ_t is (see (1))

$$x_{t+1} = \arg \min_{x \in V} \left\{ \eta_t \left\langle \sum_{s=1}^t g_s, x \right\rangle + \psi(x) \right\}.$$

842 By Theorem 6.15 in [35], this update is equivalent to

$$\begin{aligned}\tilde{x}_{t+1} &= \arg \min_{x \in \mathbb{R}^d} \left\{ \eta_t \left\langle \sum_{s=1}^t g_s, x \right\rangle + \psi(x) \right\}, \\ x_{t+1} &= \arg \min_{x \in \mathcal{B}_p} D_\psi(x, \tilde{x}_{t+1}).\end{aligned}$$

843 Now by Theorem 6.13 of [35], the first minimisation (over \mathbb{R}^d) is equivalent to

$$\tilde{x}_{t+1} = \nabla \psi^\star \left(-\eta_t \sum_{s=1}^t g_s \right),$$

844 where ψ^\star is the Fenchel conjugate of ψ .

845 For an arbitrary norm $\|\cdot\|$, the Fenchel conjugate of $f(x) = \frac{1}{r} \|x\|^r$ is $f^\star(x) = \frac{1}{r_\star} \|x\|_\star^{r_\star}$ (see
846 Lemma 2.2 in [23]). Therefore the Fenchel conjugate of $\psi(x) = \frac{1}{r} \|x\|_r^r$ is $\psi^\star(x) = \frac{1}{r_\star} \|x\|_{r_\star}^{r_\star}$ and $\nabla \psi^\star(x) =$
847 $\text{sign}(x) |x|^{r_\star - 1}$. Combining everything gives the result. \square

848 We now provide two lemmas pertaining to the Bregman projections of the FTRL update in Lemma B.1
849 for specific cases that will be of use in the proofs in Appendix E.

850 **Lemma B.2.** Consider $z = c \cdot w$ where w is a vector with all entries equal to 1 and $c > d^{-1/p}$ so
851 that $z \notin \mathcal{B}_p$. The Bregman projection $\arg \min_{x \in \mathcal{B}_p} D_\psi(x, z)$ with $\psi(x) = \frac{1}{r} \|x\|_r^r$ of z is $d^{-1/p} \cdot w$,
852 the rescaled version of w that has ℓ_p -norm equal to 1.

853 *Proof.* We make use of Lemma 5.4 in [5]: if f is a convex and differentiable function on \mathcal{B}_p then x is
854 a minimiser of $f(x)$ in \mathcal{B}_p if and only if $\nabla f(x)^T(y - x) \geq 0$ for all $y \in \mathcal{B}_p$. Consider

$$\begin{aligned}f(x) &= D_\psi(x, z) = \psi(x) - \psi(z) - \nabla \psi(z)^T(x - z), \\ \nabla f(x) &= \nabla \psi(x) - \nabla \psi(z), \\ [\nabla \psi(x)]_i &= \text{sign}(x_i) |x_i|^{r-1}\end{aligned}$$

855 Consider $x = d^{-1/p} \cdot w$. From the lemma mentioned above, it is enough to show that $\nabla f(x)^T(y -$
856 $x) \geq 0$ for all $y \in \mathcal{B}_p$:

$$\begin{aligned}
\nabla f(x)^T(y - x) &= (\nabla \psi(x) - \nabla \psi(z))^T(y - x) \\
&= (d^{-(r-1)/p} \cdot w - c^{r-1} \cdot w)^T(y - x) \\
&= (c^{r-1} - d^{-(r-1)/p})w^T(x - y) \\
&= (c^{r-1} - d^{-(r-1)/p})(d^{1-1/p} - \sum_{i=1}^d y_i) \\
&\geq (c^{r-1} - d^{-(r-1)/p})(d^{1-1/p} - \|y\|_1) \\
&\geq (c^{r-1} - d^{-(r-1)/p})(d^{1-1/p} - d^{1-1/p}\|y\|_p) \\
&\geq 0,
\end{aligned}$$

857 where we used that $c^{r-1} - d^{-(r-1)/r} > 0$ and $\|y\|_1 \leq d^{1-1/p}\|y\|_p \leq d^{1-1/p}$ for all $y \in \mathcal{B}_p$. \square

858 **Lemma B.3.** Consider $z = c \cdot e_1$ where e_1 is the first canonical basis vector and $|c| > 1$ so that
859 $z \notin \mathcal{B}_p$. The Bregman projection $\operatorname{argmin}_{x \in \mathcal{B}_p} D_\psi(x, z)$ with $\psi(x) = \frac{1}{r}\|x\|_r^r$ of z is $\operatorname{sign}(c) \cdot e_1$.

860 *Proof.* As in the proof of Lemma B.2, it is enough to show that $\nabla f(x)^T(y - x) \geq 0$ for all $y \in \mathcal{B}_p$,
861 with $x = \operatorname{sign}(c) \cdot e_1$, and

$$\begin{aligned}
f(x) &= D_\psi(x, z) = \psi(x) - \psi(z) - \nabla \psi(z)^T(x - z), \\
\nabla f(x) &= \nabla \psi(x) - \nabla \psi(z), \\
[\nabla \psi(x)]_i &= \operatorname{sign}(x_i)|x_i|^{r-1}. \\
\nabla f(x)^T(y - x) &= (\nabla \psi(x) - \nabla \psi(z))^T(y - x) \\
&= (\nabla \psi(\operatorname{sign}(c) \cdot e_1) - \nabla \psi(c \cdot e_1))^T(y - x) \\
&= \operatorname{sign}(c) \cdot (|c|^{r-1} - 1)e_1^T(x - y) \\
&= \operatorname{sign}(c) \cdot (|c|^{r-1} - 1)(\operatorname{sign}(c) - y_1) \\
&\geq 0,
\end{aligned}$$

862 where we used that $|c| > 1$ and $y_1 \leq 1$ for all $y \in \mathcal{B}_p$. \square

863 C Proofs for Section 2

864 C.1 Proof of Theorem 2.3

865 We follow and extend the analysis of FTRL from [35] (Section 7) which is closely related to the
 866 analysis in [30]. FTRL with uniformly convex regularisation was originally considered in [4] based
 867 on the analysis in [30]. Existence and unicity of the update can be handled along the same lines as
 868 Theorem 6.8 in [35] with uniform convexity.

869 The analysis begins with the following expression for the regret. We refer the reader to [35] for the
 870 proof.

871 **Lemma C.1.** *Lemma 7.1 of [35] Denote $F_t(x) = \psi_t(x) + \sum_{s=1}^{t-1} \ell_s(x)$ and set $x_t \in$
 872 $\arg \min_{x \in V} F_t(x)$. Consider $\psi_{T+1} = \psi_T$. Then, for any $u \in V$ we have*

$$\sum_{t=1}^T \ell_t(x_t) - \ell_t(u) \leq \psi_T(u) - \min_{x \in V} \psi_1(x) + \sum_{t=1}^T \left\{ F_t(x_t) - F_{t+1}(x_{t+1}) + \ell_t(x_t) \right\} \quad (4)$$

873 To bound the terms $F_t(x_t) - F_{t+1}(x_{t+1}) + \ell_t(x_t)$, we use the uniform convexity of the regularisers.
 874 In particular, we require the following result on uniformly convex functions, which is an extension of
 875 Corollary 7.7 of [35].

876 **Lemma C.2.** *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be closed, proper, sub-differentiable and μ -uniformly convex of degree
 877 r w.r.t. a norm $\|\cdot\|$. Let $x^* = \arg \min_{x \in \text{dom} f} f(x)$. Then for all $x \in \text{dom} f$ and $g \in \partial f(x)$, we have*

$$f(x) - f(x^*) \leq \frac{r-1}{r\mu^{1/(r-1)}} \|g\|_*^{r/(r-1)}.$$

878 *Proof.* By the uniform convexity of f , we have

$$\begin{aligned} f(x^*) &= \min_{z \in \text{dom} f} f(z) \\ &\geq \min_{z \in \text{dom} f} \left\{ f(x) + \langle g, z - x \rangle + \frac{\mu}{r} \|z - x\|^r \right\} \\ &\geq f(x) + \min_{z \in \mathbb{R}^d} \left\{ \langle g, z - x \rangle + \frac{\mu}{r} \|z - x\|^r \right\} \\ &= f(x) + \min_{z \in \mathbb{R}^d} \left\{ \langle g, z \rangle + \frac{\mu}{r} \|z\|^r \right\} \\ &= f(x) - \mu \max_{z \in \mathbb{R}^d} \left\{ \left\langle \frac{-g}{\mu}, z \right\rangle - \frac{1}{r} \|z\|^r \right\} \\ &= f(x) - \frac{\mu}{r_*} \left\| \frac{-g}{\mu} \right\|_*^{r_*} \\ &= f(x) - \mu^{1-r_*} \frac{\|g\|_*^{r_*}}{r_*} \\ &= f(x) - \frac{r-1}{r\mu^{1/(r-1)}} \|g\|_*^{r/(r-1)} \end{aligned}$$

879 where we used that the fenchel conjugate of $\frac{\|x\|^r}{r}$ is $\frac{\|x\|_*^{r_*}}{r_*}$. Rearranging gives the result. \square

880 Since ψ_t is proper, closed, differentiable and μ_t -uniformly convex of degree r_t with respect to $\|\cdot\|_t$
 881 and the losses are proper and convex, $F_t(x) + \ell_t(x) = \psi_t(x) + \sum_{s=1}^t \ell_s(x)$ is also proper, closed,
 882 sub-differentiable and μ_t -uniformly convex of degree r_t with respect to $\|\cdot\|_t$. Applying Lemma C.2
 883 to $F_t + \ell_t$, we have with $x_t^* = \arg \min_{x \in V} F_t(x) + \ell_t(x)$

$$\begin{aligned} F_t(x_t) - F_{t+1}(x_{t+1}) + \ell_t(x_t) &= \left(F_t(x_t) + \ell_t(x_t) \right) - \left(F_t(x_{t+1}) + \ell_t(x_{t+1}) \right) + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \\ &\leq \left(F_t(x_t) + \ell_t(x_t) \right) - \left(F_t(x_t^*) + \ell_t(x_t^*) \right) + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \\ &\leq \frac{r_t-1}{r_t\mu_t^{1/(r_t-1)}} \|g_t\|_{t*}^{r_t/(r_t-1)} + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}), \end{aligned}$$

where we used that $g_t \in \partial(F_t + \ell_t)(x_t)$ since $g_t \in \partial\ell_t(x_t)$ and $x_t = \arg \min_{x \in V} F_t(x)$. We omit some technical details but the steps from [35] extend to our setting. Plugging the above into (4) gives Theorem 2.3.

C.2 Proof of Corollary 2.4

Since ψ is μ -uniformly convex function on V of degree r with respect to $\|\cdot\|$, then the regulariser used by FTRL in round t , $\psi_t = \frac{1}{\eta_{t-1}}\psi$ is $\frac{\mu}{\eta_{t-1}}$ -uniformly convex function on V of degree r with respect to $\|\cdot\|$. Since $\eta_t \leq \eta_{t-1}$, $\psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) = \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right)\psi(x_{t+1}) \leq 0$. By the Lipschitz condition on the losses, we have $\|g_t\|_* \leq L_{\|\cdot\|}$. Applying Theorem 2.3, we have

$$\begin{aligned} \sum_{t=1}^T \ell_t(x_t) - \ell_t(u) &\leq \frac{\psi(u)}{\eta_{T-1}} + \frac{L_{\|\cdot\|}^{r_*}}{r_* \mu^{r_*-1}} \sum_{t=1}^T \eta_{t-1}^{r_*-1} \\ &\leq \frac{L_{\|\cdot\|} D^{1-1/r_*} (r_* - 1)^{1/r_*} T^{1/r_*}}{\mu^{1/r}} + \frac{L_{\|\cdot\|}^{r_*}}{r_* \mu^{r_*-1}} \sum_{t=1}^T \left(\frac{D^{1/r_*} \mu^{1/r}}{L_{\|\cdot\|} (r_* - 1)^{1/r_*} t^{1/r_*}} \right)^{r_*-1} \\ &\leq \frac{L_{\|\cdot\|} D^{1/r} (r_* - 1)^{1/r_*} T^{1/r_*}}{\mu^{1/r}} + \frac{L_{\|\cdot\|} D^{1/r}}{r_* (r_* - 1)^{1/r} \mu^{(r_*-1)(1-1/r)}} \sum_{t=1}^T \frac{1}{t^{1/r}} \\ &= \frac{L_{\|\cdot\|} D^{1/r}}{\mu^{1/r}} \left((r_* - 1)^{1/r_*} T^{1/r_*} + \frac{1}{r_* (r_* - 1)^{1/r}} \sum_{t=1}^T \frac{1}{t^{1/r}} \right). \end{aligned}$$

Now note that

$$\begin{aligned} \sum_{t=1}^T \frac{1}{t^{1/r}} &\leq \int_0^T \frac{1}{x^{1/r}} dx = \left[\frac{1}{1-1/r} x^{1-1/r} \right]_0^T = r_* T^{1/r_*} \\ \Rightarrow \sum_{t=1}^T \ell_t(x_t) - \ell_t(u) &\leq \frac{L_{\|\cdot\|} D^{1/r} T^{1/r_*}}{\mu^{1/r}} \left((r_* - 1)^{1/r_*} + \frac{1}{(r_* - 1)^{1/r}} \right). \end{aligned}$$

The proof is concluded by noting that $(r_* - 1)^{1/r_*} + \frac{1}{(r_* - 1)^{1/r}} = r^{1/r} r_*^{1/r_*}$. Lemma C.4 was helpful in finding the optimal step-size.

C.3 Proof of Theorem 2.5

We have $d \leq T$. Let $k = \lfloor T/d \rfloor \geq 1$. Let $Y_{i,j}$ be i.i.d. Rademacher random variables for $1 \leq i \leq d$, $1 \leq j \leq k$, i.e. $\mathbb{P}(Y_{i,j} = 1) = \mathbb{P}(Y_{i,j} = -1) = 1/2$. Let e_1, \dots, e_d be the canonical basis of \mathbb{R}^d . Define $g_t = LY_{i,j} \cdot e_i$ where $t = k(i-1) + j$ (for k rounds we stick to the same coordinate and draw i.i.d. Rademacher random variables). Denote the point played by \mathcal{A} by x_t and fix the loss to be $\tilde{\ell}_t(x) = g_t^T x$ (for $t > dk$, fix $\tilde{\ell}_t(x) = 0$). The subgradient is g_t , which is bounded by L in ℓ_{p_*} -norm. The point x_t depends on the losses up to time $t-1$ but not on $\tilde{\ell}_t$ and is independent of Y_t , so for all t :

$$\mathbb{E}[\tilde{\ell}_t(x_t)] = \mathbb{E}[Y_t L \cdot e_t^T x_t] = \mathbb{E}[Y_t] L e_t^T \mathbb{E}[x_t] = 0 \Rightarrow \mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_t(x_t)\right] = 0.$$

On the other hand, $u = -d^{-1/p} \sum_{i=1}^d \text{sign}\left\{\sum_{j=1}^k Y_{i,j}\right\} e_i \in \mathcal{B}_p$ gives

$$\begin{aligned} \min_{x \in \mathcal{B}_p} \sum_{t=1}^T \tilde{\ell}_t(x) &\leq \sum_{t=1}^T \tilde{\ell}_t(u) \\ &= -L d^{-1/p} \left(\sum_{i=1}^d \text{sign}\left\{\sum_{j=1}^k Y_{i,j}\right\} e_i \right)^T \left(\sum_{i=1}^d \sum_{j=1}^k Y_{i,j} e_i \right) \\ &= -L d^{-1/p} \sum_{i=1}^d \left| \sum_{j=1}^k Y_{i,j} \right|. \end{aligned}$$

903 We now make use of a result from [13] (proof of Lemma 7.2): fix $B > 0$, consider $X = \sum_{i=1}^B t_i R_i$
 904 where t_i are positive integers such that $\sum_{i=1}^B t_i = k$ and R_i are i.i.d Rademacher random variables.
 905 Then $\mathbb{E}[|X|] \geq k/\sqrt{3B}$.

906 In our case, with $B = k$ and $t_i = 1$ for all i , we have that $\mathbb{E}[|\sum_{j=1}^k Y_{i,j}|] \geq \sqrt{k/3} \geq \sqrt{T/6d}$ (since
 907 $k = \lfloor T/d \rfloor \geq T/2d$ for $T \geq d$) which gives

$$\mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_t(x_t) - \min_{x \in \mathcal{B}_p} \sum_{t=1}^T \tilde{\ell}_t(x)\right] \geq 0 + Ld^{-1/p} \sum_{i=1}^d \sqrt{\frac{T}{6d}} = Ld^{-1/p} \sqrt{\frac{Td}{6}} = L\sqrt{\frac{Td^{1-2/p}}{6}}$$

908 The result follows by: $\sup_{\ell_1, \dots, \ell_T} R_T \geq \mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_t(x_t) - \min_{x \in \mathcal{B}_p} \sum_{t=1}^T \tilde{\ell}_t(x)\right] \geq L\sqrt{\frac{Td^{1-2/p}}{6}}$.

909 C.4 Proof of Theorem 2.7

910 We have $d > T$. For $t \in \{1, \dots, T\}$, let Y_t be i.i.d. Rademacher random variables, i.e. $\mathbb{P}(Y_t = 1) =$
 911 $\mathbb{P}(Y_t = -1) = 1/2$. Let e_1, \dots, e_d be the canonical basis of \mathbb{R}^d . At time-step t , denote the point
 912 played by \mathcal{A} by x_t and fix the loss to be $\tilde{\ell}_t(x) = Y_t L e_t^T x$. The subgradient is $Y_t L e_t$, which is
 913 bounded by L in ℓ_{p^*} -norm. The point x_t depends on the losses up to time $t - 1$ but not on $\tilde{\ell}_t$ and is
 914 independent of Y_t , so for all t :

$$\mathbb{E}[\tilde{\ell}_t(x_t)] = \mathbb{E}[Y_t L e_t^T x_t] = \mathbb{E}[Y_t] L e_t^T \mathbb{E}[x_t] = 0 \implies \mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_t(x_t)\right] = 0.$$

915 On the other hand,

$$\min_{x \in \mathcal{B}_p} \sum_{t=1}^T \tilde{\ell}_t(x) = L \min_{x \in \mathcal{B}_p} x^T \left(\sum_{t=1}^T Y_t e_t \right)$$

916 is attained at $x = -T^{-1/p} \sum_{t=1}^T Y_t e_t \in \mathcal{B}_p$, giving

$$\min_{x \in \mathcal{B}_p} \sum_{t=1}^T \tilde{\ell}_t(x) = -LT^{-1/p} \sum_{t,t'=1}^T Y_t Y_{t'} e_t^T e_{t'} = -LT^{-1/p} \sum_{t=1}^T Y_t^2 = -LT^{1-1/p} = -LT^{1/p^*}.$$

917 The result follows by: $\sup_{\ell_1, \dots, \ell_T} R_T \geq \mathbb{E}\left[\sum_{t=1}^T \tilde{\ell}_t(x_t) - \min_{x \in \mathcal{B}_p} \sum_{t=1}^T \tilde{\ell}_t(x)\right] = LT^{1/p^*}$.

918 C.5 Uniform Convexity of p -OMD's regulariser

919 In this section, we provide the proof of Proposition 2.6 on the μ -uniform convexity of degree p of
 920 $\psi(x) = \frac{1}{p} \|x\|_p^p$ on \mathcal{B}_p for $p > 2$.

921 Consider $x, y \in \mathcal{B}_p$. Following the steps in Remark 2.1 of [45], using convexity of ψ we have for
 922 $\lambda \in [0, 1/2]$,

$$\begin{aligned} \psi(\lambda x + (1-\lambda)y) &= \psi\left(2\lambda\left(\frac{x+y}{2}\right) + (1-2\lambda)y\right) \\ &\leq 2\lambda\psi\left(\frac{x+y}{2}\right) + (1-2\lambda)\psi(y) \\ &= \frac{2\lambda}{p} \left\| \frac{x+y}{2} \right\|_p^p + \frac{(1-2\lambda)}{p} \|y\|_p^p. \end{aligned}$$

923 From Clarkson's inequality (equation (2.1) in [2]), we have that

$$\left\| \frac{x+y}{2} \right\|_p^p + \left\| \frac{x-y}{2} \right\|_p^p \leq \frac{\|x\|_p^p}{2} + \frac{\|y\|_p^p}{2}.$$

924 Using this in the above we have

$$\begin{aligned}
\psi(\lambda x + (1 - \lambda)y) &\leq \frac{2\lambda}{p} \frac{\|x\|_p^p}{2} + \frac{2\lambda}{p} \frac{\|y\|_p^p}{2} - \frac{2\lambda}{p} \left\| \frac{x - y}{2} \right\|_p^p + \frac{(1 - 2\lambda)}{p} \|y\|_p^p \\
&= \lambda \frac{\|x\|_p^p}{p} + (1 - \lambda) \frac{\|y\|_p^p}{p} - \frac{2\lambda}{p} \left\| \frac{x - y}{2} \right\|_p^p \\
&\leq \lambda \psi(x) + (1 - \lambda) \psi(y) - \frac{2\lambda(1 - \lambda)}{p} \left\| \frac{x - y}{2} \right\|_p^p. \tag{5}
\end{aligned}$$

925 This is an alternative characterisation of uniform convexity, we now show (following steps in
926 Definition 3.2 of [23]) that it is equivalent to our original one (Definition 2.2). From the convexity
927 and differentiability of ψ ,

$$\begin{aligned}
\psi(y) + \lambda \langle \nabla \psi(y), x - y \rangle &= \psi(y) + \langle \nabla \psi(y), [y + \lambda(x - y)] - y \rangle \\
&\leq \psi(y + \lambda(x - y)) \\
&\leq \lambda \psi(x) + (1 - \lambda) \psi(y) - \frac{2\lambda(1 - \lambda)}{p} \left\| \frac{x - y}{2} \right\|_p^p.
\end{aligned}$$

928 Rearranging,

$$\begin{aligned}
&\implies \lambda \langle \nabla \psi(y), x - y \rangle \leq \lambda(\psi(x) - \psi(y)) - \frac{2\lambda(1 - \lambda)}{p} \left\| \frac{x - y}{2} \right\|_p^p \\
&\implies \langle \nabla \psi(y), x - y \rangle \leq (\psi(x) - \psi(y)) - \frac{2(1 - \lambda)}{p} \left\| \frac{x - y}{2} \right\|_p^p \\
&\implies \psi(x) \geq \psi(y) + \langle \nabla \psi(y), x - y \rangle + \frac{2}{p} \left\| \frac{x - y}{2} \right\|_p^p,
\end{aligned}$$

929 as $\lambda \rightarrow 0$. So for any $x, y \in \mathcal{B}_p$ we have the condition of uniform convexity with $\mu = 2^{1-p}$. \square

930 **Remark C.3.** It is not possible to get the parameter of uniform convexity $\mu = 1$. Consider the
931 1-dimensional case, $x = 1, y = -1$:

$$\begin{aligned}
\psi(x) + \langle \nabla \psi(x), y - x \rangle + \frac{\mu}{p} \|x - y\|_p^p &= \frac{1}{p} + (y - x) + \frac{\mu}{p} (1 + 1)^p \\
&= \frac{1}{p} + (-1 - 1) + \frac{\mu 2^p}{p} \\
&= \frac{1 - 2p + \mu 2^p}{p}.
\end{aligned}$$

932 This is less or equal than $\psi(y) = \frac{1}{p}$ when

$$\frac{1 - 2p + \mu 2^p}{p} \leq \frac{1}{p} \implies 1 - 2p + \mu 2^p \leq 1 \implies \mu \leq p 2^{1-p}.$$

933 So our constant may be loose by a factor of p but $\mu = 1$ is not possible since $p 2^{1-p} < 1$ as soon as
934 $p > 2$.

935 In fact, we can slightly improve μ from $\frac{1}{2^{p-1}}$ to $\frac{1}{2^{p-1}-1}$ (we present our results with $\frac{1}{2^{p-1}}$ because it
936 only changes the results by a small constant and slightly avoids clutter). Here is how: In the first step
937 of the proof, we used convexity of ψ to obtain the following bound,

$$\psi\left(2\lambda\left(\frac{x+y}{2}\right) + (1 - 2\lambda)y\right) \leq 2\lambda\psi\left(\frac{x+y}{2}\right) + (1 - 2\lambda)\psi(y).$$

938 However, from (5), we have that

$$\psi(\lambda x + (1 - \lambda)y) \leq \lambda \psi(x) + (1 - \lambda) \psi(y) - \frac{2\lambda}{p} \left\| \frac{x - y}{2} \right\|_p^p, \tag{6}$$

939 and this provides a tighter bound than just using convexity:

$$\begin{aligned}
\psi\left(2\lambda\left(\frac{x+y}{2}\right) + (1 - 2\lambda)y\right) &\leq 2\lambda\psi\left(\frac{x+y}{2}\right) + (1 - 2\lambda)\psi(y) - \frac{2 \cdot 2\lambda}{p} \left\| \frac{(x+y)/2 - y}{2} \right\|_p^p \\
&\leq \lambda \psi(x) + (1 - \lambda) \psi(y) - \frac{2\lambda}{p} \left\| \frac{x - y}{2} \right\|_p^p (1 + 2^{1-p}),
\end{aligned}$$

940 where we followed similar steps as in the original proof (Clarkson's inequality). This provides an
 941 even tighter bound than (6) and applying these tighter bounds recursively gives

$$\psi\left(2\lambda\left(\frac{x+y}{2}\right) + (1-2\lambda)y\right) \leq \lambda\psi(x) + (1-\lambda)\psi(y) - \frac{2\lambda}{p} \left\|\frac{x-y}{2}\right\|_p^p \cdot \frac{1}{1-2^{1-p}},$$

942 using that $\sum_{t=0}^{\infty} (2^{1-p})^t = 1/(1-2^{1-p})$. Following the same steps for the remainder of the proof
 943 gives uniform convexity of ψ with $\mu = \frac{1}{2^{p-1}-1}$.

944 C.6 Helper lemma

945 **Lemma C.4.** Fix $a, b > 0, n > 1$. Let $f(x) = \frac{a}{x} + bx^{n-1}$ for $x > 0$. Then f is minimised at
 946 $x^* = (a/b(n-1))^{1/n}$ and

$$f(x^*) = a^{1-1/n} b^{1/n} \left(\frac{n}{n-1}\right)^{(n-1)/n} n^{1/n}.$$

947 *Proof.* Setting the derivative of f to 0 and solving gives

$$-\frac{a}{x^2} + (n-1)bx^{n-2} = 0 \implies x^* = \left(\frac{a}{(n-1)b}\right)^{1/n}.$$

948 Plugging into f gives

$$\begin{aligned} f(x^*) &= a \cdot \left(\frac{(n-1)b}{a}\right)^{1/n} + b \cdot \left(\frac{a}{(n-1)b}\right)^{(n-1)/n} \\ &= a^{1-1/n} (n-1)^{1/n} b^{1/n} + b^{1-1+1/n} a^{1-1/n} (n-1)^{1/n-1} \\ &= a^{1-1/n} b^{1/n} (n-1)^{1/n} \left(1 + \frac{1}{n-1}\right) \\ &= a^{1-1/n} b^{1/n} (n-1)^{1/n} \frac{n}{n-1} \\ &= a^{1-1/n} b^{1/n} \left(\frac{n}{n-1}\right)^{(n-1)/n} n^{1/n}. \end{aligned}$$

949

□

950 D Proofs for Section 3

951 D.1 Proof of Theorem 3.1

952 If $T \leq t_0$, then we have FTRL with fixed regulariser ϕ_p and from Corollary 2.4 we have $R_T \leq$
 953 $L(2p_*T)^{1/p_*}$ as in Section 2.3. If $T > t_0$, Theorem 2.3 gives

$$\begin{aligned} R_T &\leq \psi_T(u) - \min_{x \in V} \psi_1(x) + \sum_{t=1}^T \left\{ \frac{(r_t - 1)}{r_t \mu_t^{1/(r_t-1)}} \|g_t\|_{t*}^{r_t/(r_t-1)} + \psi_t(x_{t+1}) - \psi_{t+1}(x_{t+1}) \right\} \\ &\leq \frac{\phi_2(u)}{\eta_{T-1}} - \min_{x \in V} \phi_p(x) + \sum_{t=1}^{t_0} \left\{ 2 \frac{\eta_{t-1}^{p_*-1}}{p_*} \|g_t\|_{p_*}^{p_*} \right\} + \sum_{t=1}^{t_0-1} \left\{ \phi_p(x_{t+1}) \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t} \right) \right\} \\ &\quad + \frac{\phi_p(x_{t_0+1})}{\eta_{t_0-1}} - \frac{\phi_2(x_{t_0+1})}{\eta_{t_0}} + \sum_{t=t_0+1}^T \left\{ \frac{\eta_{t-1}}{2} \|g_t\|_2^2 + \phi_2(x_{t+1}) \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t} \right) \right\} \\ &\leq \frac{\sup_{x \in \mathcal{B}_p} \phi_p(x)}{\eta_{t_0-1}} + \sum_{t=1}^{t_0} \left\{ 2 \frac{\eta_{t-1}^{p_*-1}}{p_*} \|g_t\|_{p_*}^{p_*} \right\} + \frac{\phi_2(u)}{\eta_{T-1}} + \sum_{t=t_0+1}^T \left\{ \frac{\eta_{t-1}}{2} \|g_t\|_2^2 \right\}. \end{aligned}$$

954 The first two terms correspond to the regret of FTRL with fixed ϕ_p regularisation on t_0 rounds.
 955 Substituting the values of η_{t-1} and some algebra gives (see similar steps in the proof of Corollary 2.4)

$$\frac{\sup_{x \in \mathcal{B}_p} \phi_p(x)}{\eta_{t_0-1}} + \sum_{t=1}^{t_0} \left\{ 2 \frac{\eta_{t-1}^{p_*-1}}{p_*} \|g_t\|_{p_*}^{p_*} \right\} \leq L(2p_*t_0)^{1/p_*}.$$

956 The last two terms correspond to the regret of FTRL with fixed ϕ_2 regularisation over the remaining
 957 $T - t_0$ rounds.

$$\begin{aligned} \frac{\phi_2(u)}{\eta_{T-1}} + \sum_{t=t_0+1}^T \left\{ \frac{\eta_{t-1}}{2} \|g_t\|_2^2 \right\} &= \frac{L\sqrt{d^{1-2/p}T}}{\sqrt{2}} + \frac{L\sqrt{d^{1-2/p}}}{2\sqrt{2}} \sum_{t=t_0+1}^T \frac{1}{\sqrt{t}} \\ &\leq \frac{L\sqrt{d^{1-2/p}T}}{\sqrt{2}} + \frac{L\sqrt{d^{1-2/p}T}}{\sqrt{2}} - \frac{L\sqrt{d^{1-2/p}t_0}}{\sqrt{2}} \\ &= L\sqrt{2d^{1-2/p}T} - L\sqrt{d^{1-2/p}t_0/2}, \end{aligned}$$

958 where we used that $\sum_{t=t_0+1}^T \frac{1}{\sqrt{t}} \leq \int_{t_0}^T \frac{1}{\sqrt{x}} dx = \left[2\sqrt{x} \right]_{t_0}^T = 2(\sqrt{T} - \sqrt{t_0})$. Combining, we have

$$R_T \leq L\sqrt{2d^{1-2/p}T} + L(2p_*t_0)^{1/p_*} - L\sqrt{d^{1-2/p}t_0/2}.$$

959 The proof is concluded by $t_0 = 3^{-2p/(p-2)}d$ guaranteeing $(2p_*t_0)^{1/p_*} - \sqrt{d^{1-2/p}t_0/2} < 0$ since

$$(2p_*t_0)^{1/p_*} \leq \frac{3}{\sqrt{2}} t_0^{1/p_*} = 3t_0^{\frac{p-1}{p}-\frac{1}{2}} \sqrt{t_0/2} = 3t_0^{\frac{p-2}{2p}} \sqrt{t_0/2} = \sqrt{d^{1-2/p}t_0/2}.$$

E Proofs for Section 4

E.1 Loss construction for proofs

Many of the proofs in this section share the same loss construction, which we describe here. Assume that T is divisible by 4 (use $T - 1$, $T - 2$ or $T - 3$ if not). We define the following linear losses $\ell_t(x) = L \cdot x^T g_t$ where $g_t \in \mathcal{B}_{p^*}$ is defined as

$$g_t = \begin{cases} (-1)^t \cdot e_1, & t \leq \frac{T}{2}, \\ -v, & t > \frac{T}{2}, \end{cases}$$

where $v \in \mathcal{B}_{p^*}$ is a vector with equal entries defined as $v_{t,i} = d^{-1/p^*}$ (so that $\|v\|_{p^*} = 1$). Note that $\|v\|_p = d^{1/p-1/p^*}$. The cumulative loss of the competitor:

$$\sum_{t=1}^T \ell_t(x) = \frac{LT}{2} x^T v \implies \min_{x \in \mathcal{B}_p} \sum_{t=1}^T \ell_t(x) = -\frac{LT}{2} \frac{v^T v}{\|v\|_p} = -\frac{LT}{2} \frac{d^{1-2/p^*}}{d^{1/p-1/p^*}} = -\frac{LT}{2}. \quad (7)$$

The cumulative sum of sub-gradients used in the FTRL update:

$$L \sum_{s=1}^{t-1} g_s = L \cdot \begin{cases} -e_1, & \text{if } t \leq \frac{T}{2} \text{ is even,} \\ 0, & \text{if } t \leq \frac{T}{2} \text{ is odd,} \\ -\left(t - 1 - \frac{T}{2}\right) \cdot v, & \text{if } t > \frac{T}{2}. \end{cases} \quad (8)$$

E.2 Proofs of Proposition 4.1 and Proposition 4.5

The two propositions are special cases of the following proposition.

Proposition E.1. For $r \in [2, p]$, define $\phi_r(x) = \frac{1}{r} \|x\|_r^r$. There exists a sequence of linear L -Lipschitz losses (in ℓ_p -norm) for which FTRL with regulariser $\psi_t(x) = \frac{1}{\eta_{t-1}} \phi_r(x)$ and any sequence of decreasing η_{t-1} suffers regret

$$R_T \geq L \cdot \min\left(\frac{T}{8r}, \frac{d^{(r^*-p^*)/r^* p^*} T^{1/r^*}}{8}\right).$$

We now prove this proposition. The loss construction is described in Appendix E.1. From Lemma B.1,

$$x_{t+1} = \arg \min_{x \in \mathcal{B}_p} D_{\phi_r}\left(x, \text{sign}\left(-\eta_t \sum_{s=1}^t g_s\right) \left| -\eta_t \sum_{s=1}^t g_s \right|^{r^*-1}\right).$$

Define $\alpha_{t-1} = \min\{1, \eta_{t-1}\}$. Using (11), the points played by FTRL on are given by

- For $t \leq T/2$ odd: $x_t = 0$
- For $t \leq T/2$ even: $x_t = \alpha_{t-1}^{r^*-1} \cdot e_1$ by Lemma B.3.
- For $t > T/2$:

$$\begin{aligned} x_t &= \min\left(\frac{1}{\|w\|_p}, \left\{ \eta_{t-1} d^{-1/p^*} \left(t - 1 - \frac{T}{2}\right) \right\}^{r^*-1}\right) \cdot w \\ &= \min\left(1, d^{1-r^*/p^*} \left\{ \eta_{t-1} \left(t - 1 - \frac{T}{2}\right) \right\}^{r^*-1}\right) \cdot \frac{v}{\|v\|_p} \end{aligned}$$

by Lemma B.2 where w is a vector with equal entries equal to 1.

Fix $\eta = \eta_{T/2-1}$, $\alpha = \min\{1, \eta\}$. Using that $\eta_{t-1} \geq \eta_t$, the loss in the first half of the rounds is lower bounded as

$$\sum_{t=1}^{T/2} \ell_t(x_t) = \sum_{k=1}^{T/4} \ell_{2k}(x_{2k}) = \sum_{k=1}^{T/4} \alpha_{2k-1}^{r^*-1} e_1^T x_{2k} = \sum_{k=1}^{T/4} \alpha_{2k-1}^{r^*-1} e_1^T e_1 \geq \frac{\alpha^{r^*-1} T}{4}.$$

981 If $\alpha \geq 1$, we have $R_T \geq \frac{T}{8} \geq \frac{T}{4r}$ and we are done. So for the rest we assume that $\alpha = \eta \leq 1/2$.
 982 Let $k^* = \lfloor d^{(r_*/p_*-1)/(r_*-1)}/\eta \rfloor$, $m = \min(k^*, T/2 - 1)$. Note that $v^T v = d^{1-2/p_*} = \|v\|_p$. The
 983 losses in the second half is lower-bounded as

$$\begin{aligned} \sum_{t=T/2+1}^T \ell_t(x_t) &= - \sum_{t=T/2+1}^T x_t^T v \geq - \sum_{k=1}^{T/2-1} \min\left\{1, d^{1-r_*/p_*}(\eta k)^{r_*-1}\right\} \cdot \frac{v^T v}{\|v\|_p} \\ &= - \sum_{k=1}^{T/2-1} \min\left\{1, d^{1-r_*/p_*}(\eta k)^{r_*-1}\right\} \\ &= -d^{1-r_*/p_*} \sum_{k=1}^m (\eta k)^{r_*-1} - \left(\frac{T}{2} - 1 - m\right). \end{aligned}$$

984 We bound the sum with an integral as follows,

$$\sum_{k=1}^m k^{r_*-1} \leq \int_0^m (x+1)^{r_*-1} dx = \frac{1}{r_*} \left[(x+1)^{r_*} \right]_0^m \leq \frac{1}{r_*} (m+1)^{r_*}.$$

985 We get

$$\sum_{t=T/2+1}^T \ell_t(x_t) \geq -d^{1-r_*/p_*} \frac{\eta^{r_*-1}}{r_*} (m+1)^{r_*} - \left(\frac{T}{2} - 1 - m\right).$$

986 Using the cumulative loss of the competitor from (10), the regret is

$$\begin{aligned} R_T &\geq \frac{T}{2} + \frac{\eta^{r_*-1}T}{4} - d^{1-r_*/p_*} \frac{\eta^{r_*-1}}{r_*} (m+1)^{r_*} - \left(\frac{T}{2} - 1 - m\right) \\ &= \frac{\eta^{r_*-1}T}{4} - d^{1-r_*/p_*} \frac{\eta^{r_*-1}}{r_*} (m+1)^{r_*} + (1+m). \end{aligned}$$

987 Let's consider two cases:

988 • $k^* \geq T/2 - 1$: $m = T/2 - 1$. By the definition of k^* :

$$\begin{aligned} \frac{d^{(r_*/p_*-1)/(r_*-1)}}{\eta} &\geq \left\lfloor \frac{d^{(r_*/p_*-1)/(r_*-1)}}{\eta} \right\rfloor = k^* \geq \frac{T}{2} - 1 \implies \frac{\eta}{d^{(r_*/p_*-1)/(r_*-1)}} \left(\frac{T}{2} - 1\right) \leq 1 \\ &\implies \frac{\eta}{d^{(r_*/p_*-1)/(r_*-1)}} \frac{T}{2} \leq 1 + \frac{\eta}{d^{(r_*/p_*-1)/(r_*-1)}} \leq \frac{3}{2}, \end{aligned}$$

989 since $\eta \leq 1/2$ and $d^{(r_*/p_*-1)/(r_*-1)} \geq 1$ (recall $r \leq p$). Using this in the regret, we get

$$\begin{aligned} R_T &\geq -d^{1-r_*/p_*} \frac{\eta^{r_*-1}}{r_*} \left(\frac{T}{2}\right)^{r_*} + \frac{T}{2} \\ &= -\frac{1}{r_*} \left(\frac{\eta}{d^{(r_*/p_*-1)/(r_*-1)}} \frac{T}{2} \right)^{r_*-1} \frac{T}{2} + \frac{T}{2} \\ &\geq -\frac{1}{r_*} \left(\frac{3}{2}\right)^{r_*-1} \frac{T}{2} + \frac{T}{2} \\ &= \frac{T}{2} \left(1 - \frac{1}{r_*} \left(\frac{3}{2}\right)^{r_*-1}\right) \\ &\geq \frac{T}{4} \left(1 - \frac{1}{r_*}\right) = \frac{T}{4r_*}, \end{aligned}$$

990 where we used that $r_* \in [1, 2]$ and that $f(x) = 1 - \frac{(3/2)^{x-1}}{x} \geq \frac{1}{2}(1 - 1/x)$ for $x \in [1, 2]$.

991 • $k^* < T/2 - 1$: $m = k^*$. By the definition of k^* :

$$\begin{aligned} \frac{d^{(r_*/p_*-1)/(r_*-1)}}{\eta} &\geq \left\lfloor \frac{d^{(r_*/p_*-1)/(r_*-1)}}{\eta} \right\rfloor = k^* \implies \frac{\eta}{d^{(r_*/p_*-1)/(r_*-1)}} k^* \leq 1 \\ &\implies \frac{\eta}{d^{(r_*/p_*-1)/(r_*-1)}} (k^* + 1) \leq 1 + \frac{\eta}{d^{(r_*/p_*-1)/(r_*-1)}} \leq \frac{3}{2}, \end{aligned}$$

992 since again $\eta \leq 1/2$ and $d^{(r_*/p_*-1)/(r_*-1)} \geq 1$. We also have $k^* + 1 \geq \frac{d^{(r_*/p_*-1)/(r_*-1)}}{\eta}$. Using
 993 this in the regret, we get

$$\begin{aligned}
 R_T &= \frac{\eta^{r_*-1}T}{4} - d^{1-r_*/p_*} \frac{\eta^{r_*-1}}{r_*} (k^* + 1)^{r_*} + (1 + k^*) \\
 &= \frac{\eta^{r_*-1}T}{4} - \frac{k^* + 1}{r_*} \left(\frac{\eta}{d^{(r_*/p_*-1)/(r_*-1)}} (k^* + 1) \right)^{r_*-1} + (1 + k^*) \\
 &\geq \frac{\eta^{r_*-1}T}{4} - \frac{1 + k^*}{r_*} \left(\frac{3}{2} \right)^{r_*-1} + (1 + k^*) \\
 &= \frac{\eta^{r_*-1}T}{4} + (1 + k^*) \left(1 - \frac{1}{r_*} \left(\frac{3}{2} \right)^{r_*-1} \right) \\
 &\geq \frac{\eta^{r_*-1}T}{4} + \frac{(1 + k^*)}{2} \left(1 - \frac{1}{r_*} \right) \\
 &= \frac{\eta^{r_*-1}T}{4} + \frac{d^{(r_*/p_*-1)/(r_*-1)}}{2r_*\eta} \\
 &\geq \frac{r_*^{1/r_*}}{2^{1/r_*+2/r_*}} d^{(r_*/p_*-1)/r_*} T^{1/r_*} \geq \frac{d^{(r_*-p_*)/r_*p_*} T^{1/r_*}}{4}
 \end{aligned}$$

994 where again we used that $1 - \frac{1}{r_*} \left(\frac{3}{2} \right)^{r_*-1} \geq \frac{1}{2} (1 - 1/r_*)$ since $r_* \in [1, 2]$ and in the lstar step we
 995 minimised over η using Lemma C.4.

996 Combining both cases, we have that $R_T \geq \min\left(\frac{T}{4r}, \frac{d^{(r_*-p_*)/r_*p_*} T^{1/r_*}}{4}\right)$. If T is not divisible
 997 by 4 and we use $T - 1$, $T - 2$ or $T - 3$, we have $R_T \geq \min\left(\frac{T-3}{4r}, \frac{d^{(r_*-p_*)/r_*p_*} (T-3)^{1/r_*}}{4}\right) \geq$
 998 $\min\left(\frac{T}{8r}, \frac{d^{(r_*-p_*)/r_*p_*} (T-3)^{1/r_*}}{8}\right)$ for $T \geq 6$, concluding the proof.

999 E.3 Proof of Lemma 4.3

1000 Consider $a \in \arg \min_{z \in \mathbb{R}} \psi(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_d)$. Since ψ is sign-invariant, $-a$ is also in
 1001 the argmin. Consider $g(z) = \psi(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_d)$. It is straightforward to show that the
 1002 strong-convexity of ψ applies g . By convexity, we have

$$g(0) = g\left(\frac{1}{2}a + \frac{1}{2}(-a)\right) \leq \frac{1}{2}g(a) + \frac{1}{2}g(-a) = g(a) = \min_{z \in \mathbb{R}} g(z),$$

1003 and by strong convexity, it is actually the unique minimiser. Hence

$$\begin{aligned}
 0 = \arg \min_{z \in \mathbb{R}} \psi(\dots, x_{i-1}, z, x_{i+1}, \dots) &\implies \frac{\partial \psi(x)}{\partial x_i} \Big|_{x_i=0} = 0 \\
 &\implies \nabla \psi(x)^T e_i = 0 \quad \text{for any } x \in \mathcal{B}_p \text{ s.t. } x_i = 0. \quad (9)
 \end{aligned}$$

1004 For a set S and a vector x , denote x_{-S} the vector x with the coordinates in S replaced by 0. Denote
 1005 $S_n = \{1, \dots, n\}$. We prove the following claim by induction on $n \leq d$:

$$\psi(x) \geq \psi(x_{-S_n}) + \frac{\mu}{2} \sum_{i=1}^n x_i^2.$$

1006 **Base Case:**, by strong convexity

$$\begin{aligned}
 \psi(x) &\geq \psi(x_{-\{1\}}) + \langle \nabla \psi(x_{-\{1\}}), x - x_{-\{1\}} \rangle + \frac{\mu}{2} \|x - x_{-\{1\}}\|^2 \\
 &= \psi(x_{-\{1\}}) + \langle \nabla \psi(x_{-\{1\}}), x_1 e_1 \rangle + \frac{x_1^2 \mu}{2} \\
 &= \psi(x_{-\{1\}}) + \frac{x_1^2 \mu}{2} \quad \text{using (9)}.
 \end{aligned}$$

1007 **Inductive Step:** suppose true for k . Similarly to the base case: by strong convexity,

$$\begin{aligned}\psi(x_{-S_n}) &\geq \psi(x_{-S_{n+1}}) + \langle \nabla \psi(x_{-S_{n+1}}), x_{-S_n} - x_{-S_{n+1}} \rangle + \frac{\mu}{2} \|x_{-S_n} - x_{-S_{n+1}}\|^2 \\ &= \psi(x_{-S_{n+1}}) + x_{n+1} \langle \nabla \psi(x_{-S_{n+1}}), e_{n+1} \rangle + \frac{\mu}{2} x_{n+1}^2 \\ &= \psi(x_{-S_{n+1}}) + \frac{\mu}{2} x_{n+1}^2 \quad \text{using (9)}.\end{aligned}$$

1008 The result follows by the inductive hypothesis:

$$\begin{aligned}\psi(x) &\geq \psi(x_{-S_n}) + \frac{\mu}{2} \sum_{i=1}^n x_i^2 \\ &\geq \psi(x_{-S_{n+1}}) + \frac{\mu}{2} x_{n+1}^2 + \frac{\mu}{2} \sum_{i=1}^n x_i^2 \\ &\geq \psi(x_{-S_{n+1}}) + \frac{\mu}{2} \sum_{i=1}^{n+1} x_i^2.\end{aligned}$$

1009 When $n = d$, we have $\psi(x) \geq \frac{\mu}{2} \|x\|_2^2$.

1010 E.4 Proof of Lemma 4.4

1011 As discussed in Remark 4.9, we prove a more general version of Lemma 4.4 for coordinate-wise
1012 step-sizes, where the FTRL update is allowed to have a different step-size $\eta_{t-1,i}$ for each coordinate:

$$1013 \quad x_t = \arg \min_{x \in V} \left\{ \psi(x) + \sum_{i=1}^d \eta_{t-1,i} \cdot x_i \sum_{s=1}^{t-1} g_{s,i} \right\}.$$

1014 We consider a slight variation of the loss construction described in Appendix E.1: Assume that T is
1015 divisible by 4 (use $T-1$, $T-2$ or $T-3$ if not). We define the following linear losses $\ell_t(x) = L \cdot x^T g_t$
1016 where $g_t \in \mathcal{B}_{p^*}$ is defined as

$$g_t = \begin{cases} (-1)^t \cdot v, & t \leq 2 \\ (-1)^t \cdot e_{i(t)}, & 2 < t \leq \frac{T}{2}, \\ -v, & t > \frac{T}{2}, \end{cases}$$

1017 where $v \in \mathcal{B}_{p^*}$ is a vector with equal entries defined as $v_{t,i} = d^{-1/p^*}$ (so that $\|v\|_{p^*} = 1$) and
1018 $i(t) = \arg \max_{i \in [d]} \eta_{2\lfloor (t-1)/2 \rfloor, i}$ (i.e. the coordinate of the largest step-size in the previous even
1019 round). Note that $\|v\|_p = d^{1/p-1/p^*}$. The cumulative loss of the competitor:

$$\sum_{t=1}^T \ell_t(x) = \frac{LT}{2} x^T v \implies \min_{x \in \mathcal{B}_p} \sum_{t=1}^T \ell_t(x) = -\frac{LT}{2} \frac{v^T v}{\|v\|_p} = -\frac{LT}{2} \frac{d^{1-2/p^*}}{d^{1/p-1/p^*}} = -\frac{LT}{2}. \quad (10)$$

1020 The cumulative sum of sub-gradients used in the FTRL update:

$$L \sum_{s=1}^{t-1} g_s = L \cdot \begin{cases} -e_{i(t)}, & \text{if } t \leq \frac{T}{2} \text{ is even,} \\ 0, & \text{if } t \leq \frac{T}{2} \text{ is odd,} \\ -\left(t-1-\frac{T}{2}\right) \cdot v, & \text{if } t > \frac{T}{2}. \end{cases} \quad (11)$$

1021 • First we consider $2 < t \leq T/2$. When t is odd, $x_t = \arg \min_{x \in \mathcal{B}_p} \psi(x) = 0$. When t is
1022 even,

$$\begin{aligned}x_t &= \arg \min_{x \in \mathcal{B}_p} \left\{ \psi(x) - \eta_{t-1,i(t)} L e_{i(t)}^T x \right\} \\ &\implies \psi(x_t) - \eta_{t-1,i(t)} L x_t^T e_{i(t)} \leq \psi(e_{i(t)}) - \eta_{t-1,i(t)} L e_{i(t)}^T e_{i(t)} = 1 - L \eta_{t-1,i(t)} \\ &\implies \ell_t(x_t) = x_t^T e_{i(t)} \geq 1 - \frac{1}{\eta_{t-1,i(t)} L} \geq 1 - \frac{1}{\eta_{t-1,i(T/2)} L} \quad \text{by def of } i(t) \\ &\geq 1 - \frac{1}{\eta_{T/2,i(T/2)} L} \geq 1/2,\end{aligned}$$

1023 when $\eta := \eta_{T/2, i(T/2)} = \max_{i \in [d]} \eta_{T/2, i} \geq 2/L$. So we have $(-1$ accounts for first 2
 1024 rounds not being like the rest)

$$\sum_{t=1}^{T/2} g_t^T x_t \geq \begin{cases} T/4 - 1, & \text{if } \eta \geq 2/L \\ 0, & \text{if } \eta < 2/L. \end{cases}$$

1025 Hence if $\eta \geq 2/L$, we have $R_T \geq LT/4 - 1$ and the statement of the theorem holds. If
 1026 $\eta < 2/L$, we look to the second half of the rounds.

1027 • Let's now consider $t > T/2$ and assume $\eta < 2/L$. Note that by definition of η , we have
 1028 the for all $t \geq T/2$ and for all $i \in [d]$, $\eta_{t, i} \leq \eta \leq 2/L$. Fix $\beta_t = t - T/2 - 1$. The FTRL
 1029 update is

$$x_t = \arg \min_{x \in \mathcal{B}_p} \left\{ \psi(x) - L\beta_t \cdot x^T (\eta_{t-1} \odot v) \right\}.$$

1030 Let $u = v/\|v\|_p$ be the competitor. We can write $x_t = \lambda_t u + \alpha_t u^\perp$ ($\lambda_t > 0$) as a component
 1031 in the direction of u and a component orthogonal to u . We have

$$\psi(x_t) \geq \frac{\mu}{2} \|x_t\|_2^2 = \frac{\mu}{2} (\lambda_t^2 \|u\|_2^2 + \alpha_t^2 \|u^\perp\|_2^2) \geq \frac{1}{2} \lambda_t^2 \mu d^{1-2/p}.$$

1032 Now from the FTRL update, (in the first implication, we use that $\eta_{t-1, i} \leq \eta$ and $x_{t, i} \geq$
 1033 $0, v_i \geq 0$)

$$\begin{aligned} \psi(x_t) - L\beta_t x_t^T (\eta_{t-1} \odot v) &\leq 0 \implies \frac{1}{2} \lambda_t^2 \mu d^{1-2/p} \leq \eta L \beta_t x_t^T v \\ &\implies \frac{1}{2} \lambda_t^2 \mu d^{1-2/p} \leq \eta L \beta_t \lambda_t \\ &\implies \lambda_t \leq \frac{2\eta L \beta_t}{\mu d^{1-2/p}} \\ &\implies \ell_t(x_t) = -L \cdot v^T x_t = -L \lambda_t \geq -L \frac{2\eta L \beta_t}{\mu d^{1-2/p}} \geq -L \frac{4\beta_t}{\mu d^{1-2/p}}, \end{aligned}$$

1034 since $\eta \leq 2/L$. If $d \geq (4T/\mu)^{p/(p-2)}$, we have for all $t \leq T$

$$\begin{aligned} \ell_t(x_t) &\geq -L \frac{4\beta_t}{\mu d^{1-2/p}} \geq -L \frac{\beta_t}{T} \geq -\frac{L}{2} \\ &\implies R_T \geq \frac{LT}{2} + \sum_{t=T/2+1}^T \ell_t(x_t) \geq \frac{LT}{2} - \frac{LT}{4} = \frac{LT}{4}. \end{aligned}$$

1035 If T is not divisible by 4 and we use $T-1, T-2$ or $T-3$, we have $R_T \geq \frac{L(T-3)}{4} - 1 \geq \frac{LT}{8}$ for
 1036 $T \geq 6 + \frac{8}{L}$, concluding the proof.

1037 E.5 Proof of Lemma 4.7

1038 Assume there exists a constant $c > 0$ such that for all T and any sequence of losses, $R_T \leq cL\sqrt{T}$.

1039 Consider $T > 16c^2$ and a multiple of 4. We define the following linear losses $\ell_t(x) = x \cdot g_t$ where
 1040 $g_t \in [-1, 1]$ is defined as

$$g_t = \begin{cases} (-1)^t \cdot L, & t \leq \frac{T}{2}, \\ -L, & t > \frac{T}{2}, \end{cases}$$

1041 Recall that the FTRL update is $x_t = \arg \min_{x \in [-1, 1]} \left\{ \eta_{t-1} \left(\sum_{s=1}^{t-1} g_s \right) \cdot x + \psi(x) \right\}$. Set $\eta = \eta_{T/2-1}$.
 1042 With this sequence of losses, the points played by FTRL satisfy

1043 • for $t \leq T/2 + 1$ and t odd, we have $\sum_{s=1}^{t-1} g_s = 0$, so $x_t = 0$.

1044 • for $t \leq T/2 + 1$ and t even, we have $\sum_{s=1}^{t-1} g_s = -L$ so $x_t = \arg \min_{x \in [-1,1]} \{-\eta_{t-1}x +$
 1045 $\psi(x)\}$. For $t < t' \leq T/2$ (both even), we have

$$\begin{aligned}
 -\eta_{t'-1}Lx_{t'} + \psi(x_{t'}) &\leq -\eta_{t'-1}Lx_t + \psi(x_t) && \text{using the definition of } x_{t'} \\
 &= -\eta_{t-1}Lx_t + \psi(x_t) + L(\eta_{t-1} - \eta_{t'-1})x_t \\
 &\leq -\eta_{t-1}Lx_{t'} + \psi(x_{t'}) + L(\eta_{t-1} - \eta_{t'-1})x_t && \text{using the definition of } x_t \\
 \implies (\eta_{t-1} - \eta_{t'-1})Lx_{t'} &\leq (\eta_{t-1} - \eta_{t'-1})Lx_t \\
 \implies x_{t'} &\leq x_t && \text{using that } \eta_{t'-1} \leq \eta_{t-1}.
 \end{aligned}$$

1046 So for all $t \leq T/2$ even, we have $x_t \geq x_{T/2}$.

1047 • for $t > T/2$, we have $\sum_{s=1}^{t-1} g_s = -L(t - T/2 - 1)$ so $x_t = \arg \min_{x \in [-1,1]} \{-\eta_{t-1}L(t -$
 1048 $T/2 - 1) \cdot x + \psi(x)\}$.

1049 The regret can then be written as follows

$$R_T = \sum_{t=1}^T \ell_t(x_t) - \left(-\frac{LT}{2}\right) \geq \frac{LT}{2} + \frac{LT}{4}x_{T/2} - L \sum_{t=T/2+1}^T x_t, \quad (12)$$

1050 from which we can show the series of following statements.

1051 1. We first show $\max_{2 \leq t \leq \lceil 2c\sqrt{T} \rceil} x_{\frac{T}{2}+t} \geq \frac{1}{2}$: if not, $x_{\frac{T}{2}+t} < \frac{1}{2}$ for all $t \leq \lceil 2c\sqrt{T} \rceil$ and from
 1052 (12):

$$\begin{aligned}
 R_T &\geq \frac{LT}{2} - L \left(\sum_{t=T/2+1}^{T/2+\lceil 2c\sqrt{T} \rceil} x_t + \sum_{t=T/2+\lceil 2c\sqrt{T} \rceil+1}^T x_t \right) \\
 &> \frac{LT}{2} - L \left(\sum_{t=T/2+1}^{T/2+\lceil 2c\sqrt{T} \rceil} \frac{1}{2} + \sum_{t=T/2+\lceil 2c\sqrt{T} \rceil+1}^T 1 \right) \\
 &= \frac{LT}{2} - \frac{L}{2} \lceil 2c\sqrt{T} \rceil - L \left(T - \lceil 2c\sqrt{T} \rceil - \frac{T}{2} \right) \\
 &\geq \frac{L}{2} \lceil 2c\sqrt{T} \rceil \\
 &> cL\sqrt{T},
 \end{aligned}$$

1053 which contradicts our initial assumption that $R_T \leq cL\sqrt{T}$ so we must have
 1054 $\max_{2 \leq t \leq \lceil 2c\sqrt{T} \rceil} x_{\frac{T}{2}+t} \geq \frac{1}{2}$. Note that $2c\sqrt{T} < T/2$ is ensured by $T > 16c^2$.

1055 2. Next, we show that $\eta \geq \frac{\psi(1/2)}{2cL\sqrt{T}}$: let $t^* = \arg \max_{2 \leq t \leq \lceil 2c\sqrt{T} \rceil} x_{\frac{T}{2}+t}$, by the definition of
 1056 $x_{\frac{T}{2}+t^*}$:

$$\begin{aligned}
 0 &\geq -\eta_{\frac{T}{2}+t^*-1}L \left(\frac{T}{2} + t^* - 1 - \frac{T}{2} \right) x_{\frac{T}{2}+t^*} + \psi(x_{\frac{T}{2}+t^*}) \\
 \implies \eta_{\frac{T}{2}+t^*-1}L(t^* - 1) &\geq \psi(1/2) \\
 \implies \eta = \eta_{T/2-1} &\geq \eta_{\frac{T}{2}+t^*-1} \geq \frac{\psi(1/2)}{L(t^* - 1)} \geq \frac{\psi(1/2)}{2cL\sqrt{T}}.
 \end{aligned}$$

1057 where in the first implication, we used that $\psi(x_{\frac{T}{2}+t^*}) \geq \psi(1/2)$ (since ψ is increasing on
 1058 $[0, 1]$ and $x_{\frac{T}{2}+t^*} \geq 1/2$) and $x_{\frac{T}{2}+t^*} \leq 1$.

1059 3. From (12), we also have $R_T \geq \frac{LT}{4}x_{T/2}$. To achieve $R_T \leq cL\sqrt{T}$, we must have
 1060 $x_{T/2} \leq \frac{4c}{\sqrt{T}}$.

1061
1062

4. By the definition of $x_{T/2} = \arg \min_{x \in [-1, 1]} \{-\eta Lx + \psi(x)\}$, for any $x \in [4c/\sqrt{T}, 1]$ we have

$$\begin{aligned} -\eta Lx_{T/2} + \psi(x_{T/2}) &\leq -\eta Lx + \psi(x) \\ \implies \psi(x) &\geq \eta L(x - x_{T/2}) \geq \eta L\left(x - \frac{4c}{\sqrt{T}}\right) \geq \frac{\psi(1/2)}{2c\sqrt{T}}\left(x - \frac{4c}{\sqrt{T}}\right) \\ \implies \psi\left(\frac{5c}{\sqrt{T}}\right) &\geq \frac{\psi(1/2)}{2c\sqrt{T}} \frac{c}{\sqrt{T}} = \frac{\psi(1/2)}{2T}. \end{aligned}$$

1063
1064

5. Now fix $x \in [0, 1]$. There exists T (multiple of 4) such that $x \in \left[\frac{5c}{\sqrt{T+4}}, \frac{5c}{\sqrt{T}}\right]$. Using that ψ is increasing on $[0, 1]$ and from the previous point, we have

$$\psi(x) \geq \psi\left(\frac{5c}{\sqrt{T+4}}\right) \geq \frac{\psi(1/2)}{2(T+4)} \geq \frac{\psi(1/2)}{2(T+4)} \frac{T}{25c^2} x^2 \geq \frac{\psi(1/2)}{100c^2} x^2,$$

1065

using that $T/(T+4) \geq 1/2$ for $T \geq 8$. The result is shown with $\mu = \psi(1/2)/100c^2$.

F Proofs of Section 5

Throughout this section, we will use the notation R_T for the pseudo-regret. In fact since a randomized learner is equivalent to a random choice of deterministic learners, we will consider in the proofs below deterministic learners and the regret is equal to the pseudo-regret. In addition, since our loss constructions are oblivious to the learner's actions, even for randomised learners, the pseudo-regret is equal to the regret.

We split the proof into the case where $p > 4/3$ is "large" and the case where $p \in [1, 4/3]$ is "small" and consider separate loss constructions for each. We first highlight the intuition of the loss constructions.

- **For $p > 4/3$,** we take inspiration from the loss construction which [6] use to prove a $\Omega(d\sqrt{T})$ lower bound for low-dimensional ($d^2 < T$) ℓ_p -balls with $p > 2$. The construction consists of linear Gaussian losses where the mean of each coordinate is the same distance from 0 but the learner does not know the sign. When the dimension is large enough, the learner does not acquire enough information to determine the signs of these means in the T rounds to get sub-linear regret. **This construction will not work when $p \leq 4/3$** because when p is close to 1, the lack of a distinct corner in the ℓ_p -ball allows any point on the boundary (including $\pm e_1$) with correct signs to achieve similar loss to the competitor (a corner). The learner can therefore focus on $\pm e_1$ simplifying the problem to one-dimension where sub-linear regret is achievable.
- **For $p \leq 4/3$,** the construction consists of linear Gaussian losses where the mean vector has a single non-zero positive entry, unknown to the learner. When the dimension is large enough, it does not acquire enough information to determine the non-zero coordinate of the mean in the T rounds to get sub-linear regret. **This construction will not work when $p > 4/3$** because for $p \gg 1$ the learner can exploit the ℓ_p -ball's proximity to the hypercube by playing points with all coordinates close to -1 , bypassing the need to identify the correct non-zero mean coordinate.

We present the proofs with a Lipschitz constant of 1 but they extend straightforwardly to arbitrary $L > 0$.

F.1 Case $p > 4/3$

Theorem F.1. Fix T and $\delta > 0$. Consider $p > 4/3$ and

$$d > \max\left\{16T, \frac{1}{c_1} \log \frac{C_1 T}{\delta}, \left(\frac{1}{c_1 p_*} \log \frac{C_1 T}{\delta}\right)^{p_*/2}, e^2\right\},$$

for some universal constants c_1, C_1 . For any OCO algorithm with bandit feedback on $V = \mathcal{B}_p$, there exists a sequence of random linear losses $(\ell_t)_{t \in [T]}$ with sub-gradients $(g_t)_{t \in [T]}$ such that $\|g_t\|_{p_*} \leq 1$ for all rounds t with probability at least $1 - \delta$ and

$$\mathbb{E}[R_T] \geq \frac{T}{80},$$

where the expectation is with respect to the randomness of the losses.

F.1.1 Proof

The following loss construction and analysis is inspired from the proof of Theorem 4 of [6]. Their construction is designed for the low-dimensional setting in such a way that the learner has to balance exploration and exploitation rounds. We only consider the losses corresponding to exploration rounds and generalize the analysis to the high-dimensional setting.

Let $\varepsilon > 0$ be such that $\varepsilon^{p_*} = 1/d$. Let $T < \alpha d$ (with $\alpha = 1/16$). For a fixed $\xi \in \{-1, 1\}^d$, define the losses as $\ell_t(x) = x^T g_t^\xi$ where $g_t^\xi \sim \mathcal{N}(\varepsilon \xi, \frac{1}{d^{2/p_*}} I_d)$ (i.i.d.). We show that (when ξ is sampled uniformly at the start and fixed throughout the rounds)

$$\mathbb{E}_\xi \mathbb{E}_{g_t^\xi} [R_T] \geq \frac{T}{16}.$$

We use \mathbb{E}_ξ for the expectation with respect to ξ and $\mathbb{E}_{g_t^\xi}$ for the expectation with respect to g_t^ξ with ξ fixed. We will also use $x_{t,i}$ to mean the i -th coordinate of x_t .

1108 For fixed ξ : $\mathbb{E}_{g_t^\xi}[\ell_t(x)] = \mathbb{E}[x^T g_t^\xi] = \varepsilon \cdot x^T \xi$. So the competitor $x^* = \arg \min_{x \in V} \varepsilon \cdot \xi^T x =$
 1109 $-d^{-1/p} \xi$. Let us define $\bar{x} = \frac{1}{T} \sum_{t=1}^T \mathbb{E}[x_t]$. In particular one has

$$\mathbb{E}_{g_t^\xi}[R_T] = \varepsilon T \cdot \xi^T (\bar{x} - x^*).$$

1110 The following lemma expresses the expected regret in terms of the expected number of rounds and
 1111 coordinates for which the learner plays on the wrong side of ξ . The proof is in Appendix F.1.3.

Lemma F.2 (Generalization of Lemma 6 of [6]).

$$\mathbb{E}_{g_t^\xi}[R_T] \geq \frac{\varepsilon^{p_*}}{p_*} \cdot \mathbb{E}_{g_t^\xi} \left[\sum_{t=1}^T \sum_{i=1}^d \mathbb{I}\{x_{t,i} \xi_i \geq 0\} \right].$$

1112 And now the next lemma shows that the expected number of rounds and coordinates for which the
 1113 learner plays on the wrong side of ξ is linear in both T and d . The proof is in Appendix F.1.4.

1114 **Lemma F.3.** With $T < \alpha d = \frac{1}{16}$, we have $\mathbb{E}_\xi \mathbb{E}_{g_t^\xi} \sum_{t=1}^T \sum_{i=1}^d \mathbb{I}\{x_{t,i} \xi_i \geq 0\} \geq \frac{dT}{4}$.

1115 Combining both lemmas, we have

$$\mathbb{E}[R_T] \geq \frac{1}{p_*} \varepsilon^{p_*} \cdot \mathbb{E} \left[\sum_{t=1}^T \sum_{i=1}^d \mathbb{I}\{x_{t,i} \xi_i \geq 0\} \right] \geq \frac{1}{4p_*} \varepsilon^{p_*} T d = \frac{1}{4p_*} T \geq \frac{T}{16},$$

1116 since $p > 4/3$ so $p_* \leq 4$. Now to ensure the Lipschitz-condition with high-probability, we get an
 1117 extra factor of $1/5$ (see the next section), concluding the proof.

1118 F.1.2 Bound on Sub-gradients

1119 Recall that $p > 4/3$ so $p_* \leq 4$. Fix $\xi \in \{-1, 1\}^d$. $g_t \sim \mathcal{N}(\varepsilon \xi, d^{-2/p_*} I_d)$. So $g_t = d^{-1/p_*} X + \varepsilon \xi$
 1120 where $X \sim \mathcal{N}(0, I_d)$. From [43], we have

$$\begin{aligned} \mathbb{E}[\|X\|_{p_*}] &\leq \left(\mathbb{E} \left[\sum_{i=1}^d |X_i|^{p_*} \right] \right)^{1/p_*} = \left(\mathbb{E} \left[\sum_{i=1}^d 2^{p_*/2} \frac{\Gamma((p_*+1)/2)}{\sqrt{\pi}} \right] \right)^{1/p_*} \leq \sqrt{2} \left(\frac{3}{4} \right)^{1/p_*} d^{1/p_*} \leq 2d^{1/p_*}. \\ \implies \mathbb{E}[\|g_t\|_{p_*}] &\leq 2 + \varepsilon d^{1/p_*} = 2 + 1 = 3. \end{aligned}$$

1121 Fix $\delta > 0$.

1122 • **For $p_* \leq 2$:** By Theorem 1.1 in [37] for some constants $C_1, c_1 > 0$,

$$\mathbb{P}(\|X\|_{p_*} \leq (1 + \beta) \mathbb{E}[\|X\|_{p_*}]) \geq 1 - C_1 \exp(-c_1 \beta^2 d).$$

1123 Assuming $d \geq \frac{1}{c_1} \log \frac{C_1 T}{\delta}$, we have $\beta = \sqrt{\frac{1}{c_1 d} \log \frac{C_1 T}{\delta}} \leq 1$ and

$$\mathbb{P}(\|X\|_{p_*} \leq (1 + \beta) \mathbb{E}[\|X\|_{p_*}]) \geq 1 - \frac{\delta}{T}.$$

1124 • **For $2 < p_* \leq 4$:** By Theorem 1.1 in [37] for some constants $C_1, c_1 > 0$,

$$\mathbb{P}(\|X\|_{p_*} \leq (1 + \beta) \mathbb{E}[\|X\|_{p_*}]) \geq 1 - C_1 \exp(-c_1 \beta p_* d^{2/p_*}).$$

1125 Assuming $d \geq \left(\frac{1}{c_1 p_*} \log \frac{C_1 T}{\delta} \right)^{p_*/2}$, we have $\beta = \frac{1}{c_1 q d^{2/p_*}} \log \frac{C_1 T}{\delta} \leq 1$ and

$$\mathbb{P}(\|X\|_{p_*} \leq (1 + \beta) \mathbb{E}[\|X\|_{p_*}]) \geq 1 - \frac{\delta}{T}.$$

1126 In both cases, with probability at least $1 - \delta/T$

$$\begin{aligned} \|X\|_{p_*} &\leq (1 + \beta) \mathbb{E}[\|X\|_{p_*}] \leq 2 \mathbb{E}[\|X\|_{p_*}] \leq 4d^{1/p_*}, \\ \implies \|g_t\|_{p_*} &\leq d^{-1/p_*} \|X\|_{p_*} + \varepsilon d^{1/p_*} \leq 4 + 1 = 5. \end{aligned}$$

1127 By a union bound over all rounds, with probability $1 - \delta$, $\|g_t\|_{p_*} \leq 5$ for all rounds t . So rescaling
 1128 the losses by a factor of 5 gives sub-gradients whose ℓ_{p_*} -norm is bounded by 1 with high-probability
 1129 and a regret bound of:

$$\mathbb{E}[R_T] \geq \frac{T}{80}.$$

1130 **F.1.3 Proof of Lemma F.2**

1131 Let $W_t = \{i \in [d] : x_{t,i}\xi_i < 0\}$ and $S = \mathbb{E}\left[\sum_{t=1}^T \sum_{i=1}^d \mathbb{I}\{x_{t,i}\xi_i \geq 0\}\right]$. We have that

$$\begin{aligned}\mathbb{E}_{g_t^\xi}[R_T] &= \varepsilon T \cdot \xi^T(\bar{x} - x^\star) \\ &= \varepsilon \sum_{t=1}^T \mathbb{E}_{g_t^\xi} \left[\sum_{i \notin W_t} \xi_i x_{t,i} \right] + \varepsilon \sum_{t=1}^T \mathbb{E}_{g_t^\xi} \left[\sum_{i \in W_t} \xi_i x_{t,i} \right] + \varepsilon T d^{1/p_\star} \\ &\geq \varepsilon \sum_{t=1}^T \mathbb{E}_{g_t^\xi} \left[\sum_{i \in W_t} \xi_i x_{t,i} \right] + \varepsilon T d^{1/p_\star}\end{aligned}$$

1132 Therefore, it is sufficient to show that

$$\varepsilon \sum_{t=1}^T \mathbb{E}_{g_t^\xi} \left[\sum_{i \in W_t} \xi_i x_{t,i} \right] + \varepsilon T d^{1/p_\star} \geq \frac{\varepsilon^{p_\star} S}{p_\star}.$$

1133 Since $\|x_{t,W_t}\|_p \leq 1$ (we use x_{t,W_t} to mean that the coordinates of x_t that are not in W_t are 0), by
1134 Holder's inequality we know that

$$\varepsilon \sum_{i \in W_t} \xi_i x_{t,i} = (x_{t,W_t})^T (\varepsilon \xi_{W_t}) \geq -\|x_{t,W_t}\|_p \|\varepsilon \xi_{W_t}\|_{p_\star} \geq -|W_t|^{1/p_\star} \varepsilon.$$

1135 Noting that (see (14) below)

$$|W_t|^{1/p_\star} \varepsilon = ((d - |W_t^C|) \varepsilon^{p_\star})^{1/p_\star} \leq (d \varepsilon^{p_\star})^{1/p_\star} - \frac{1}{p_\star} \varepsilon^{p_\star} |W_t^C|, \quad (13)$$

1136 we have

$$\begin{aligned}\varepsilon \sum_{i \in W_t} \xi_i x_{t,i} &\geq \frac{1}{p_\star} \varepsilon^{p_\star} |W_t^C| - (d \varepsilon^{p_\star})^{1/p_\star} = \frac{1}{p_\star} \varepsilon^{p_\star} \sum_{i=1}^d \mathbb{I}\{x_{t,i}\xi_i \geq 0\} - (d \varepsilon^{p_\star})^{1/p_\star} \\ \Rightarrow \varepsilon \sum_{t=1}^T \mathbb{E}_{g_t^\xi} \left[\sum_{i \in W_t} \xi_i x_{t,i} \right] &\geq \frac{\varepsilon^{p_\star}}{p_\star} \mathbb{E}_{g_t^\xi} \left[\sum_{t=1}^T \sum_{i=1}^d \mathbb{I}\{x_{t,i}\xi_i \geq 0\} \right] - T (d \varepsilon^{p_\star})^{1/p_\star} = \frac{\varepsilon^{p_\star} S}{p_\star} - \varepsilon T d^{1/p_\star},\end{aligned}$$

1137 which concludes the proof.

1138 **Proof of (13):** Since x^{1/p_\star} is concave: for all $x, y \in \mathbb{R}$, $x^{1/p_\star} \leq y^{1/p_\star} + \frac{1}{p_\star} y^{-1/p} (x - y)$. In
1139 particular, with $x = \varepsilon^{p_\star} (d - s)$, $y = \varepsilon^{p_\star} d$, we have

$$\varepsilon (d - s)^{1/p_\star} \leq \varepsilon d^{1/p_\star} - \frac{1}{p_\star} \frac{\varepsilon^{p_\star} s}{(\varepsilon^{p_\star} d)^{1/p}} = \varepsilon d^{1/p_\star} - \frac{1}{p_\star} \varepsilon^{p_\star} s, \quad (14)$$

1140 since $\varepsilon^{p_\star} d = 1$. Using $s = |W_t^C|$ gives the result.

1141 **F.1.4 Proof of Lemma F.3**

1142 At round t conditioned on ξ , the observed feedback is

$$f_t^\xi := x_t^T g_t^\xi \sim \mathcal{N}(\varepsilon \cdot x_t^T \xi, \sigma_t^2), \text{ where } \sigma_t^2 = \frac{\|x_t\|_2^2}{d^{2/p_\star}}.$$

1143 Denote p_ξ for the law of the observed feedback up to time T conditioned on ξ , i.e., the law of
1144 $(f_1^\xi, \dots, f_T^\xi)$. Consider ξ and ξ' differing only in coordinate $i \in d$. By Pinsker's inequality we have

$$d_{\text{TV}}(p_\xi, p_{\xi'}) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(p_\xi, p_{\xi'})}.$$

1145 By the chain rule for the KL divergence / operations on conditional densities:

$$\begin{aligned}
D_{\text{KL}}(p_\xi, p_{\xi'}) &= \mathbb{E}_{(f_1, \dots, f_T) \sim p_\xi} \left[\log \frac{p_\xi(f_1, \dots, f_T)}{p_{\xi'}(f_1, \dots, f_T)} \right] \\
&= \sum_{t=1}^T \mathbb{E}_{(f_1, \dots, f_t) \sim p_\xi} \left[\log \frac{p_\xi(f_t | f_{t-1}, \dots, f_1)}{p_{\xi'}(f_t | f_{t-1}, \dots, f_1)} \right] \\
&= \sum_{t=1}^T \mathbb{E}_{(f_1, \dots, f_{t-1}) \sim p_\xi} \left\{ \mathbb{E}_{f_t \sim p_\xi(\cdot | f_1, \dots, f_{t-1})} \left[\log \frac{p_\xi(f_t | f_{t-1}, \dots, f_1)}{p_{\xi'}(f_t | f_{t-1}, \dots, f_1)} \right] \right\}.
\end{aligned}$$

1146 Now since x_t is a deterministic function of f_1, \dots, f_{t-1} and given x_t , $f_t \sim \mathcal{N}(\varepsilon x_t^T \xi, \sigma_t^2)$ under p_ξ ,
1147 we have that the inner expectation is a KL divergence between Gaussians:

$$\begin{aligned}
\mathbb{E}_{f_t \sim p_\xi(\cdot | f_1, \dots, f_{t-1})} \left[\log \frac{p_\xi(f_t | f_{t-1}, \dots, f_1)}{p_{\xi'}(f_t | f_{t-1}, \dots, f_1)} \right] &= \frac{(\varepsilon x_t^T \xi - \varepsilon x_t^T \xi')^2}{2\sigma_t^2} = \frac{4\varepsilon^2 x_{t,i}^2}{2\sigma_t^2} = \frac{2\varepsilon^2 x_{t,i}^2}{\sigma_t^2} \\
\Rightarrow D_{\text{KL}}(p_\xi, p_{\xi'}) &= 2 \sum_{t=1}^T \mathbb{E}_{(f_1, \dots, f_{t-1}) \sim p_\xi} \left[\frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2} \right] = 2 \sum_{t=1}^T \mathbb{E}_{p_\xi} \left[\frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2} \right] \\
\Rightarrow d_{\text{TV}}(p_\xi, p_{\xi'}) &\leq \sqrt{\sum_{t=1}^T \mathbb{E}_{p_\xi} \left[\frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2} \right]}.
\end{aligned}$$

1148 Now, using ξ_{-i} to refer to all the coordinates of ξ except the i -th and ξ_{i+} (resp. ξ_{i-}) to denote that
1149 the i -th coordinate of ξ is $+1$ (resp. -1),

$$\begin{aligned}
\mathbb{E}_\xi \left[\mathbb{E}_{p_\xi} \left[\sum_{t=1}^T \mathbb{I}\{x_{t,i} \xi_i \geq 0\} \right] \right] &= \mathbb{E}_{\xi_{-i}} \mathbb{E}_{\xi_i} \left[\mathbb{E}_{p_\xi} \left[\sum_{t=1}^T \mathbb{I}\{x_{t,i} \xi_i \geq 0\} \right] | \xi_{-i} \right] \\
&= \frac{1}{2} \mathbb{E}_{\xi_{-i}} \left[\mathbb{E}_{p_{\xi_{i+}}} \left[\sum_{t=1}^T \mathbb{I}\{x_{t,i} \cdot 1 \geq 0\} \right] + \mathbb{E}_{p_{\xi_{i-}}} \left[\sum_{t=1}^T \mathbb{I}\{x_{t,i} \cdot (-1) \geq 0\} \right] \right] \\
&= \frac{1}{2} \mathbb{E}_{\xi_{-i}} \left[\mathbb{E}_{p_{\xi_{i+}}} \left[\sum_{t=1}^T \mathbb{I}\{x_{t,i} \geq 0\} \right] + \mathbb{E}_{p_{\xi_{i-}}} \left[\sum_{t=1}^T 1 - \mathbb{I}\{x_{t,i} > 0\} \right] \right] \\
&\geq \frac{T}{2} + \frac{1}{2} \sum_{t=1}^T \mathbb{E}_{\xi_{-i}} \left[\mathbb{E}_{p_{\xi_{i+}}} [\mathbb{I}\{x_{t,i} \geq 0\}] - \mathbb{E}_{p_{\xi_{i-}}} [\mathbb{I}\{x_{t,i} \geq 0\}] \right] \\
&\geq \frac{T}{2} - \frac{1}{4} \sum_{t=1}^T \mathbb{E}_{\xi_{-i}} \left[d_{\text{TV}}(p_{\xi_{i+}}, p_{\xi_{i-}}) + d_{\text{TV}}(p_{\xi_{i-}}, p_{\xi_{i+}}) \right] \quad \text{using Pinsker's inequality} \\
&= \frac{T}{2} - \frac{T}{4} \mathbb{E}_{\xi_{-i}} \left[d_{\text{TV}}(p_{\xi_{i+}}, p_{\xi_{i-}}) + d_{\text{TV}}(p_{\xi_{i-}}, p_{\xi_{i+}}) \right] \\
&\geq \frac{T}{2} - \frac{T}{4} \mathbb{E}_{\xi_{-i}} \left[\sqrt{\sum_{t=1}^T \mathbb{E}_{g_{t,i+}} \frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2}} + \sqrt{\sum_{t=1}^T \mathbb{E}_{g_{t,i-}} \frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2}} \right] \\
&= \frac{T}{2} - \frac{T}{2} \mathbb{E}_\xi \left[\sqrt{\sum_{t=1}^T \mathbb{E}_{g_t^\xi} \frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2}} \right].
\end{aligned}$$

1150 Summing over all possible coordinates i , we get:

$$\frac{1}{T} \sum_{i=1}^d \mathbb{E}_\xi \left[\mathbb{E}_{p_\xi} \left[\sum_{t=1}^T \mathbb{I}\{x_{t,i} \xi_i \geq 0\} \right] \right] \geq \frac{d}{2} - \frac{1}{2} \sum_{i=1}^d \mathbb{E}_\xi \left[\sqrt{\sum_{t=1}^T \mathbb{E}_{g_t^\xi} \frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2}} \right].$$

1151 Note that due to the concavity of the square-root:

$$\begin{aligned}
\sum_{i=1}^d \mathbb{E}_\xi \left[\sqrt{\sum_{t=1}^T \mathbb{E}_{g_t^\xi} \frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2}} \right] &\leq \sum_{i=1}^d \sqrt{\sum_{t=1}^T \mathbb{E}_{\xi, g_t^\xi} \frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2}} \\
&= d \frac{1}{d} \sum_{i=1}^d \sqrt{\sum_{t=1}^T \mathbb{E}_{\xi, g_t^\xi} \frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2}} \\
&\leq d \sqrt{\frac{1}{d} \sum_{i=1}^d \sum_{t=1}^T \mathbb{E}_{\xi, g_t^\xi} \frac{\varepsilon^2 x_{t,i}^2}{\sigma_t^2}} \\
&= \sqrt{d \sum_{t=1}^T \mathbb{E}_{\xi, g_t^\xi} \frac{\varepsilon^2 \|x_t\|_2^2}{\sigma_t^2}}.
\end{aligned}$$

1152 So we get

$$\begin{aligned}
\frac{1}{T} \mathbb{E}_\xi \mathbb{E}_{g_t^\xi} \sum_{t=1}^T \sum_{i=1}^d \mathbb{I}\{x_{t,i} \xi_i \geq 0\} &\geq \frac{d}{2} - \sqrt{d \sum_{t=1}^T \mathbb{E}_{\xi, g_t^\xi} \frac{\varepsilon^2 \|x_t\|_2^2}{\sigma_t^2}} \\
&= \frac{d}{2} - \sqrt{d^{1+2/p_\star} T \varepsilon^2} \\
&= \frac{d}{2} - \sqrt{dT} \\
&\geq d \left(\frac{1}{2} - \sqrt{\alpha} \right) \\
&= \frac{d}{4}
\end{aligned}$$

1153 since $\varepsilon = (1/d)^{1/p_\star}$ and $T < \alpha d = d/16$.

1154 **F.2 Case $1 < p \leq 4/3$**

1155 **Theorem F.4.** Fix T and $\delta > 0$. Consider $p \in (1, 4/3]$ and

$$d > \max \left\{ (128 p_\star T)^2, \left(\frac{1}{c_1 p_\star} \log \frac{C_1 T}{\delta} \right)^{p_\star/2}, e^2 \right\},$$

1156 for some universal constants c_1, C_1 . For any OCO algorithm with bandit feedback on $V = \mathcal{B}_p$, there
1157 exists a sequence of random linear losses $(\ell_t)_{t \in [T]}$ with sub-gradients $(g_t)_{t \in [T]}$ such that $\|g_t\|_{p_\star} \leq 1$
1158 for all rounds t with probability at least $1 - \delta$ and

$$\mathbb{E}[R_T] \geq \frac{T}{16},$$

1159 where the expectation is with respect to the randomness of the losses.

1160 **F.2.1 Proof**

1161 Before the start of the game, draw $Y \sim \text{Unif}(1, \dots, d)$ and define the losses as $\ell_t(x) = x^T g_t^Y$ where

$$g_{t,i}^Y \sim \begin{cases} \mathcal{N}(0, \sigma^2), & \text{if } i \neq Y, \\ \mathcal{N}(1/2, \sigma^2), & \text{if } i = Y, \end{cases}$$

1162 where $\sigma = (8\sqrt{p_\star} d^{1/p_\star})^{-1}$. We show that $\mathbb{E}_Y \mathbb{E}_{\ell_1, \dots, \ell_T} R_T \geq T/16$.

1163 Fix $\alpha \in [0, 1]$ and define $A_i(\alpha) = \{t : y_{t,i} \geq -\alpha\}$. Then

$$\begin{aligned}
\mathbb{E}[R_T|Y=i] &= \frac{T}{2} + \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^T y_{t,i}|Y=i\right] \\
&= \frac{T}{2} + \frac{1}{2}\mathbb{E}\left[\sum_{t \notin A_i(\alpha)} y_{t,i} + \sum_{t \in A_i(\alpha)} y_{t,i}|Y=i\right] \\
&\geq \frac{T}{2} + \frac{1}{2}\mathbb{E}\left[\sum_{t \notin A_i(\alpha)} (-1) + \sum_{t \in A_i(\alpha)} (-\alpha)|Y=i\right] \\
&= \frac{T}{2} - \frac{1}{2}\mathbb{E}\left[T - |A_i(\alpha)| + \alpha|A_i(\alpha)||Y=i\right] \\
&= (1-\alpha)\frac{1}{2}\mathbb{E}\left[|A_i(\alpha)||Y=i\right] \\
&= (1-\alpha)\frac{1}{2}\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\{y_{t,i} \geq -\alpha\}|Y=i\right].
\end{aligned}$$

1164 The following lemma bounds the expected number of rounds where the learner suffers large regret (as
1165 measured by α) when $Y = i$ compared to an environment where all coordinates of g_t are 0-mean for
1166 all t (i.e. there is no better direction). We denote E_{p_0} expectations with respect to this environment.
1167 The proof is in Appendix F.2.3.

1168 **Lemma F.5.** Let $\sigma_t^2 = \|y_t\|_2^2 \cdot \sigma^2$, then

$$\mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\{y_{t,i} \geq -\alpha\}|Y=i\right] \geq \mathbb{E}_{p_0}\left[\sum_{t=1}^T \mathbb{I}\{y_{t,i} \geq -\alpha\}\right] - T\sqrt{\sum_{t=1}^T \mathbb{E}_{p_0}\left[\frac{y_{t,i}^2}{8\sigma_t^2}\right]}$$

1169 From the lemma, we have

$$\mathbb{E}[R_T|Y=i] \geq (1-\alpha)\frac{1}{2}\left\{\mathbb{E}_{p_0}\left[\sum_{t=1}^T \mathbb{I}\{y_{t,i} \geq -\alpha\}\right] - T\sqrt{\sum_{t=1}^T \mathbb{E}_{p_0}\left[\frac{y_{t,i}^2}{8\sigma_t^2}\right]}\right\}.$$

1170 Taking an expectation with respect to Y we have:

$$\begin{aligned}
\mathbb{E}[R_T] &= \frac{1}{d}\sum_{i=1}^d \mathbb{E}[R_T|Y=i] \\
&\geq \frac{(1-\alpha)}{2d}\sum_{i=1}^d \left\{\mathbb{E}_{p_0}\left[\sum_{t=1}^T \mathbb{I}\{y_{t,i} \geq -\alpha\}\right] - T\sqrt{\sum_{t=1}^T \mathbb{E}_{p_0}\left[\frac{y_{t,i}^2}{8\sigma_t^2}\right]}\right\} \\
&= \frac{(1-\alpha)}{2d}\left\{\mathbb{E}_{p_0}\left[\sum_{t=1}^T \sum_{i=1}^d \mathbb{I}\{y_{t,i} \geq -\alpha\}\right] - T\sum_{i=1}^d \sqrt{\sum_{t=1}^T \mathbb{E}_{p_0}\left[\frac{y_{t,i}^2}{8\sigma_t^2}\right]}\right\}.
\end{aligned}$$

1171 For the first term: fix $\alpha = (d/2)^{-1/p}$ and note that if $\sum_{i=1}^d \mathbb{I}\{y_{t,i} \geq -\alpha\} < d/2$, then there are
1172 more than $d/2$ coordinates of y_t whose value is less than $-\alpha$, this means

$$\|y_t\|_p^p > \frac{d}{2}\alpha^p = \frac{d}{2}\frac{2}{d} = 1,$$

1173 which contradicts $y_t \in \mathcal{B}_p$ so we must have $\sum_{i=1}^d \mathbb{I}\{y_{t,i} \geq -\alpha\} \geq d/2$.

1174 For the second term, using Jensen's inequality and the concavity of \sqrt{x} ,

$$\frac{1}{d}\sum_{i=1}^d \sqrt{\sum_{t=1}^T \mathbb{E}_{p_0}\left[\frac{y_{t,i}^2}{8\sigma_t^2}\right]} \leq \sqrt{\frac{1}{d}\sum_{i=1}^d \sum_{t=1}^T \mathbb{E}_{p_0}\left[\frac{y_{t,i}^2}{8\sigma_t^2}\right]} = \sqrt{\frac{1}{d}\sum_{t=1}^T \mathbb{E}_{p_0}\left[\frac{\|y_t\|_2^2}{8\sigma_t^2}\right]} = \frac{1}{2\sigma}\sqrt{\frac{T}{2d}}.$$

1175 Combining we have

$$\mathbb{E}[R_T] \geq \frac{1-\alpha}{2} \left(\frac{T}{2} - T \frac{1}{2\sigma} \sqrt{\frac{T}{2d}} \right).$$

1176 • Since $d \geq e^2 \geq 2^{p+1}$, $1 - \alpha = 1 - (2/d)^{1/p} \geq 1 - 2^{-p/p} = 1/2$.

1177 • If $d \geq (128p_*T)^2$ then $d \geq 2T/\sigma^2$ and:

$$\frac{T}{2} - T \frac{1}{2\sigma} \sqrt{\frac{T}{2d}} \geq \frac{T}{2} - \frac{T}{4} = \frac{T}{4}.$$

1178 The condition on d follows from the definition of σ and $p_* \geq 4$:

$$d \geq (128p_*T)^2 \geq (128qT)^{1/(1-2/p_*)} \implies d^{1-2/p_*} \geq 128qT \implies d \geq \frac{128p_*T}{d^{2/p_*}} = \frac{2T}{\sigma^2}.$$

1179 We hence have $\mathbb{E}[R_T] \geq T/16$. The following section ensures that the Lipschitz-condition is
1180 satisfied with high-probability.

1181 F.2.2 Bound on sub-gradients

1182 Recall that $p \leq 4/3$, so $p_* \geq 4$. Given $Y = i$, $g_t^Y \sim \mathcal{N}(\frac{1}{2}e_i, \sigma^2 I_d)$. So $g_t = \sigma X + \frac{1}{2}e_i$ where
1183 $X \sim \mathcal{N}(0, I_d)$. From [43], we have

$$\mathbb{E}[\|X\|_{p_*}] \leq \left(\mathbb{E} \left[\sum_{i=1}^d |X_i|^{p_*} \right] \right)^{1/p_*} = \left(\mathbb{E} \left[\sum_{i=1}^d 2^{p_*/2} \frac{\Gamma((p_*+1)/2)}{\sqrt{\pi}} \right] \right)^{1/p_*}.$$

1184 From [3] (Theorem 2.2), we have

$$\begin{aligned} \Gamma\left(\frac{p_*+1}{2}\right) &= \Gamma\left(\frac{p_*-1}{2} + 1\right) \leq \sqrt{2\pi} \left(\frac{p_*-1}{2}\right)^{(p_*-1)/2} \exp\left(-\frac{p_*-1}{2}\right) \sqrt{2\frac{p_*-1}{2}} \leq 2\sqrt{\pi} p_*^{p_*/2} e^{-(p_*-1)/2} \\ \implies \left(2^{p_*/2} \frac{\Gamma((p_*+1)/2)}{\sqrt{\pi}}\right)^{1/p_*} &\leq 2\sqrt{p_*} \\ \implies \mathbb{E}[\|X\|_{p_*}] &\leq 2\sqrt{p_*} d^{1/p_*} \\ \implies \mathbb{E}[\|g_t\|_{p_*}] &\leq 2\sqrt{p_*} \sigma d^{1/p_*} + \frac{1}{2}. \end{aligned}$$

1185 Fix $\delta > 0$. By Theorem 1.1 in [37] for some constants $C_1, c_1 > 0$,

$$\mathbb{P}\left(\|X\|_{p_*} \leq (1+\beta)\mathbb{E}[\|X\|_{p_*}]\right) \geq 1 - C_1 \exp(-c_1 \beta p_* d^{2/p_*}).$$

1186 Assuming $d \geq \left(\frac{1}{c_1 p_*} \log \frac{C_1 T}{\delta}\right)^{p_*/2}$, we have $\beta = \frac{1}{c_1 q d^{2/p_*}} \log \frac{C_1 T}{\delta} \leq 1$ and

$$\mathbb{P}\left(\|X\|_{p_*} \leq (1+\beta)\mathbb{E}[\|X\|_{p_*}]\right) \geq 1 - \frac{\delta}{T}.$$

1187 So with probability at least $1 - \delta/T$

$$\begin{aligned} \|X\|_{p_*} &\leq (1+\beta)\mathbb{E}[\|X\|_{p_*}] \leq 2\mathbb{E}[\|X\|_{p_*}] \leq 4\sqrt{p_*} d^{1/p_*}, \\ \implies \|g_t\|_{p_*} &\leq 4\sqrt{p_*} \sigma d^{1/p_*} + \frac{1}{2} \leq 1, \end{aligned}$$

1188 where the final inequality follows from $\sigma \leq \frac{1}{8\sqrt{p_*} d^{1/p_*}}$

1189 By a union bound over all rounds, with probability $1 - \delta$, $\|g_t\|_{p_*} \leq 1$ for all rounds t .

1190 **F.2.3 Proof of Lemma F.5**

1191 Given $Y = i$, the observed feedback at round t is exactly

$$f_t^Y := y_t^T g_t^Y \sim \mathcal{N}(\frac{1}{2}y_{t,i}, \sigma_t^2), \text{ where } \sigma_t^2 = \|y_t\|_2^2 \sigma^2.$$

1192 Denote p_i for the law of the observed feedback up to time T given $Y = i$, i.e., the law of (f_1^i, \dots, f_T^i) .
 1193 Denote p_0 (use $Y = 0$ in notation), the law of the observed feedback up to time T when all coordinates
 1194 of g_t are 0-mean for all t (i.e. there is no better direction). Under p_0 , $f_t^0 \sim \mathcal{N}(0, \sigma_t^2)$. By the definition
 1195 of the total-variation distance (d_{TV}),

$$\begin{aligned} & \mathbb{E}[\mathbb{I}\{y_{t,i} \geq -\alpha\} | Y = 0] - \mathbb{E}[\mathbb{I}\{y_{t,i} \geq -\alpha\} | Y = i] \leq d_{\text{TV}}(p_0, p_i) \\ \implies & \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\{y_{t,i} \geq -\alpha\} | Y = 0\right] - \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\{y_{t,i} \geq -\alpha\} | Y = i\right] \leq T d_{\text{TV}}(p_0, p_i). \end{aligned}$$

1196 By Pinsker's inequality we have

$$d_{\text{TV}}(p_0, p_i) \leq \sqrt{\frac{1}{2} D_{\text{KL}}(p_0, p_i)}.$$

1197 By the chain rule for the KL divergence / operations on conditional densities:

$$\begin{aligned} D_{\text{KL}}(p_0, p_i) &= \mathbb{E}_{(f_1, \dots, f_T) \sim p_0} \left[\log \frac{p_0(f_1, \dots, f_T)}{p_i(f_1, \dots, f_T)} \right] \\ &= \sum_{t=1}^T \mathbb{E}_{(f_1, \dots, f_t) \sim p_0} \left[\log \frac{p_0(f_t | f_{t-1}, \dots, f_1)}{p_i(f_t | f_{t-1}, \dots, f_1)} \right] \\ &= \sum_{t=1}^T \mathbb{E}_{(f_1, \dots, f_{t-1}) \sim p_0} \left\{ \mathbb{E}_{f_t \sim p_0(\cdot | f_1, \dots, f_{t-1})} \left[\log \frac{p_0(f_t | f_{t-1}, \dots, f_1)}{p_i(f_t | f_{t-1}, \dots, f_1)} \right] \right\}. \end{aligned}$$

1198 Now since y_t is a deterministic function of f_1, \dots, f_{t-1} and given y_t , $f_t \sim \mathcal{N}(0, \sigma_t^2)$ under p_0
 1199 and $f_t \sim \mathcal{N}(\frac{1}{2}y_{t,i}, \sigma_t^2)$ under p_i , we have that the inner expectation is a KL divergence between
 1200 Gaussians:

$$\begin{aligned} & \mathbb{E}_{f_t \sim p_0(\cdot | f_1, \dots, f_{t-1})} \left[\log \frac{p_0(f_t | f_{t-1}, \dots, f_1)}{p_i(f_t | f_{t-1}, \dots, f_1)} \right] = \frac{(0 - y_{t,i}/2)^2}{2\sigma_t^2} = \frac{y_{t,i}^2}{8\sigma_t^2} \\ \implies & D_{\text{KL}}(p_0, p_i) = \sum_{t=1}^T \mathbb{E}_{(f_1, \dots, f_{t-1}) \sim p_0} \left[\frac{y_{t,i}^2}{8\sigma_t^2} \right] = \sum_{t=1}^T \mathbb{E}_{p_0} \left[\frac{y_{t,i}^2}{8\sigma_t^2} \right] \\ \implies & d_{\text{TV}}(p_0, p_i) \leq \sqrt{\sum_{t=1}^T \mathbb{E}_{p_0} \left[\frac{y_{t,i}^2}{8\sigma_t^2} \right]}. \end{aligned}$$

1201 Combining gives the result.

1202 **F.3 Case $p \rightarrow 1$**

1203 **Theorem F.6.** Fix T and $\delta > 0$. Consider $d > \max\{T/\delta, 8^4 e^4 T^2, 8\}$ and $p \in [1, 1 + 1/\log d]$. For
 1204 any OCO algorithm with bandit feedback on $V = \mathcal{B}_p$, there exists a sequence of random linear losses
 1205 $(\ell_t)_{t \in [T]}$ with sub-gradients $(g_t)_{t \in [T]}$ such that $\|g_t\|_{p^*} \leq 1$ for all rounds t with probability at least
 1206 $1 - \delta$ and

$$\mathbb{E}[R_T] \geq \frac{T}{16},$$

1207 where the expectation is with respect to the randomness of the losses.

1208 **F.3.1 Proof**

1209 We use almost the same loss construction as the proof of Theorem F.4. Before the start of the game,
 1210 draw $Y \sim \text{Unif}(1, \dots, d)$ and define the losses as $\ell_t(x) = x^T g_t^Y$ where

$$g_{t,i}^Y \sim \begin{cases} \mathcal{N}(0, \sigma^2), & \text{if } i \neq Y, \\ \mathcal{N}(1/2, \sigma^2), & \text{if } i = Y, \end{cases}$$

1211 where $\sigma = (4\sqrt{2} \exp(1) \sqrt{\log d})^{-1}$. The only difference with the proof of Theorem F.4 being the
 1212 value of σ . We also follow the same steps until we reach for $\alpha = (2/d)^{1/p}$:

$$\mathbb{E}[R_T] \geq \frac{1-\alpha}{2} \left(\frac{T}{2} - T \frac{1}{2\sigma} \sqrt{\frac{T}{2d}} \right).$$

- 1213 • From $d \geq 8$, we have $p \leq 2$ and $2^{p+1} \leq 8 \leq d$ so $1-\alpha = 1 - (2/d)^{1/p} \geq 1 - 2^{-p/p} = 1/2$.
- 1214 • From the definition of $\sigma = (4\sqrt{2} \exp(1) \sqrt{\log d})^{-1}$

$$\frac{1}{2\sigma} \sqrt{\frac{T}{2d}} = 2 \exp(1) \sqrt{\frac{T \log d}{d}} \leq 2 \exp(1) \sqrt{\frac{T \sqrt{d}}{d}} = 2 \exp(1) \sqrt{\frac{T}{\sqrt{d}}} \leq \frac{1}{4}$$

1215 since $d \geq 8^4 e^4 T^2$. This gives:

$$\frac{T}{2} - T \frac{1}{2\sigma} \sqrt{\frac{T}{2d}} \geq \frac{T}{2} - \frac{T}{4} = \frac{T}{4}.$$

1216 We hence have $\mathbb{E}[R_T] \geq T/16$. The following section ensures that the Lipschitz-condition is
 1217 satisfied with high-probability.

1218 **F.3.2 Bound on sub-gradients**

1219 Recall that $p \leq 1 + \frac{1}{\log d}$ so $p_* \geq \frac{\log d}{1} + 1$. We have

$$\mathbb{E}[\|X\|_\infty] \leq \sqrt{2 \log d}.$$

1220 By the Borell-TIS inequality,

$$\mathbb{P}(\|X\|_\infty > \mathbb{E}[\|X\|_\infty] + \beta) \leq \exp(-\beta^2/2).$$

1221 So with probability $1 - \delta/T$, we have (using $d > T/\delta$)

$$\begin{aligned} \|X\|_\infty &\leq \mathbb{E}[\|X\|_\infty] + \sqrt{2 \log \frac{T}{\delta}} \leq \sqrt{2 \log d} + \sqrt{2 \log \frac{T}{\delta}} \leq 2\sqrt{2 \log d} \\ \implies \|g_t\|_{p_*} &\leq \sigma \|X\|_{p_*} + \frac{1}{2} \leq \sigma d^{1/p_*} \|X\|_\infty + \frac{1}{2} \leq \sigma \exp(1) 2\sqrt{2 \log d} + \frac{1}{2} = 1 \end{aligned}$$

1222 where the final equality follows from $\sigma = (4\sqrt{2} \exp(1) \sqrt{\log d})^{-1}$ and we also used that $d^{1/p_*} \leq e$.

1223 By a union bound over all rounds, with probability $1 - \delta$, $\|g_t\|_{p_*} \leq 1$ for all rounds t .

1224 G Results for Online Mirror Descent (OMD)

1225 G.1 OMD with uniformly-convex regularisation

1226 The results in this section are from [39] (Proposition 7). We include them for completeness.

1227 Let $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a proper, closed and differentiable μ -uniformly convex function³ on V of degree
1228 $r > 2$ w.r.t. a norm $\|\cdot\|$. The Bregman Divergence w.r.t. ψ is defined for all $x, y \in \mathbb{R}^d$ as

$$D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle.$$

1229 Given $x_1 \in V$, at time-step $t = 1, \dots, T$, Online Mirror Descent (OMD) with step-size $\eta_t > 0$ outputs
1230 the following update where $g_t \in \partial \ell_t(x_t)$,

$$x_{t+1} = \arg \min_{x \in V} \left\{ \eta_t \langle g_t, x \rangle + D_\psi(x, x_t) \right\}. \quad (15)$$

1231 The standard regret bound of OMD stems from the following one-step regret bound lemma (e.g. see
1232 Lemma 6.9 in [35]).

1233 **Lemma G.1.** *The iterates (15) of OMD satisfy for all $u \in V$,*

$$\ell_t(x_t) - \ell_t(u) \leq \langle g_t, x_t - x_{t+1} \rangle + \frac{D_\psi(u, x_t) - D_\psi(u, x_{t+1}) - D_\psi(x_{t+1}, x_t)}{\eta_t}$$

1234 From Lemma G.1 and the uniform convexity of ψ , we can bound the regret of OMD.

1235 **Theorem G.2.** *The iterates (15) of OMD with decreasing step-size $\eta_{t+1} \leq \eta_t$ ($1 \leq t \leq T$) satisfy*
1236 *for all $u \in V$ (recall that r_\star is the conjugate of r , i.e. $1/r + 1/r_\star = 1$),*

$$\sum_{t=1}^T \ell_t(x_t) - \ell_t(u) \leq \max_{1 \leq t \leq T} \frac{D_\psi(u, x_t)}{\eta_T} + \frac{1}{r_\star \mu^{r_\star-1}} \sum_{t=1}^T \eta_t^{r_\star-1} \|g_t\|_\star^{r_\star}. \quad (16)$$

1237 *If the step-sizes are constant: $\eta_t = \eta$ ($1 \leq t \leq T$), we have*

$$\sum_{t=1}^T \ell_t(x_t) - \ell_t(u) \leq \frac{D_\psi(u, x_1)}{\eta} + \frac{\eta^{r_\star-1}}{r_\star \mu^{r_\star-1}} \sum_{t=1}^T \|g_t\|_\star^{r_\star}. \quad (17)$$

1238 *Proof.* By the uniform convexity of ψ , $D_\psi(x_{t+1}, x_t) \geq \frac{\mu}{r} \|x_t - x_{t+1}\|^r$. Using this in Lemma G.1
1239 along with Hölder's inequality, we have for all $u \in V$,

$$\begin{aligned} \ell_t(x_t) - \ell_t(u) &\leq \langle g_t, x_t - x_{t+1} \rangle + \frac{D_\psi(u, x_t) - D_\psi(u, x_{t+1})}{\eta_t} - \frac{\mu}{r} \frac{\|x_t - x_{t+1}\|^r}{\eta_t} \\ &\leq \|g_t\|_\star \|x_t - x_{t+1}\| + \frac{D_\psi(u, x_t) - D_\psi(u, x_{t+1})}{\eta_t} - \frac{\mu}{r} \frac{\|x_t - x_{t+1}\|^r}{\eta_t}. \end{aligned}$$

1240 Consider $f(x) = \frac{1}{r} |x|^r$. Then the Fenchel conjugate of f is $f^\star(y) = \frac{1}{r_\star} |y|^{r_\star}$ (see Lemma 2.2 in
1241 [23]) and from Fenchel's inequality, we have $xy \leq \frac{1}{r} |x|^r + \frac{1}{r_\star} |y|^{r_\star}$, which we use in the following,

$$\begin{aligned} \|g_t\|_\star \|x_t - x_{t+1}\| &= \left(\frac{\eta_t^{1/r}}{\mu^{1/r}} \|g_t\|_\star \right) \cdot \left(\frac{\mu^{1/r}}{\eta_t^{1/r}} \|x_t - x_{t+1}\| \right) \\ &\leq \frac{1}{r_\star} \left(\frac{\eta_t^{1/r}}{\mu^{1/r}} \|g_t\|_\star \right)^{r_\star} + \frac{1}{r} \left(\frac{\mu^{1/r}}{\eta_t^{1/r}} \|x_t - x_{t+1}\| \right)^r \\ &= \frac{\eta_t^{r_\star-1}}{r_\star \mu^{r_\star-1}} \|g_t\|_\star^{r_\star} + \frac{\mu}{r} \frac{\|x_t - x_{t+1}\|^r}{\eta_t}, \end{aligned}$$

³The function ψ can be defined on a subset $X \subseteq \mathbb{R}^d$ but conditions on its behaviour on the boundary of X are then required for OMD to be well defined. For simplicity, we consider ψ defined on \mathbb{R}^d , though the results in this section hold more generally (see Theorem 6.7 of [35] for more detail).

1242 where we used that $r = r_*/(r_* - 1)$. Plugging this into the above inequality,

$$\ell_t(x_t) - \ell_t(u) \leq \frac{\eta_t^{r_*-1}}{r_*\mu^{r_*-1}} \|g_t\|_*^{r_*} + \frac{D_\psi(u, x_t) - D_\psi(u, x_{t+1})}{\eta_t}. \quad (18)$$

1243 Denoting $D = \max_{1 \leq t \leq T} D_\psi(u, x_t)$, the result follows by summing t over all rounds,

$$\begin{aligned} \sum_{t=1}^T \ell_t(x_t) - \ell_t(u) &\leq \sum_{t=1}^T \left(\frac{D_\psi(u, x_t)}{\eta_t} - \frac{D_\psi(u, x_{t+1})}{\eta_t} \right) + \frac{1}{r_*\mu^{r_*-1}} \sum_{t=1}^T \eta_t^{r_*-1} \|g_t\|_*^{r_*} \\ &= \frac{D_\psi(u, x_1)}{\eta_1} - \frac{D_\psi(u, x_{T+1})}{\eta_T} + \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) D_\psi(u, x_{t+1}) + \frac{1}{r_*\mu^{r_*-1}} \sum_{t=1}^T \eta_t^{r_*-1} \|g_t\|_*^{r_*} \\ &\leq \frac{D}{\eta_1} + D \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) + \frac{1}{r_*\mu^{r_*-1}} \sum_{t=1}^T \eta_t^{r_*-1} \|g_t\|_*^{r_*} \\ &= \frac{D}{\eta_1} + D \left(\frac{1}{\eta_T} - \frac{1}{\eta_1} \right) + \frac{1}{r_*\mu^{r_*-1}} \sum_{t=1}^T \eta_t^{r_*-1} \|g_t\|_*^{r_*} \\ &= \frac{D}{\eta_T} + \frac{1}{r_*\mu^{r_*-1}} \sum_{t=1}^T \eta_t^{r_*-1} \|g_t\|_*^{r_*}. \end{aligned}$$

1244 For constant step-size, the result follows similarly by summing (18) over t , giving a telescoping sum,

$$\begin{aligned} \sum_{t=1}^T \ell_t(x_t) - \ell_t(u) &\leq \frac{D_\psi(u, x_1) - D_\psi(u, x_{T+1})}{\eta} + \frac{\eta^{r_*-1}}{r_*\mu^{r_*-1}} \sum_{t=1}^T \|g_t\|_*^{r_*} \\ &\leq \frac{D_\psi(u, x_1)}{\eta} + \frac{\eta^{r_*-1}}{r_*\mu^{r_*-1}} \sum_{t=1}^T \|g_t\|_*^{r_*}, \end{aligned}$$

1245 which concludes the proof. \square

1246 G.1.1 Regret bounds

1247 When we have L -Lipschitz losses w.r.t. $\|\cdot\|$ and we can bound $D_\psi(u, x_1) < D$, then the regret bound
1248 (17) for constant step-sizes becomes

$$R_T \leq \frac{D}{\eta} + \eta^{r_*-1} \frac{TL^{r_*}}{r_*\mu^{r_*-1}}.$$

1249 Assuming the time-horizon T is known, optimising the above bound w.r.t. η using Lemma C.4 gives

$$R_T \leq \frac{r^{1/r}}{\mu^{1/r}} LD^{1/r} T^{1/r_*}, \quad (19)$$

1250 for $\eta = (Dp/T)^{1/r_*} \mu^{1/r} / L$. With $r = 2$, we recover the standard regret bound of OMD using a
1251 strongly-convex regulariser $R_T \leq L\sqrt{2DT/\mu}$.

1252 G.1.2 Anytime and adaptive bounds

1253 When T is unknown, we can use the bound in (16) and the time-varying step-size

$$\eta_t = \frac{D_{\max}^{1/r_*} (r-1)^{1/r_*} \mu^{1/r}}{L} \cdot \frac{1}{t^{1/r_*}} \quad \text{to get} \quad R_T \leq \frac{LD_{\max}^{1/r} r^{1/r} r_*^{1/r_*} T^{1/r_*}}{\mu^{1/r}}, \quad (20)$$

1254 where D_{\max} is a bound on $\max_{1 \leq t \leq T} D_\psi(u, x_t)$. Though this can be unbounded when D is bounded,
1255 for our purposes of ℓ_p -balls, D_{\max} will only be a constant away from D . The doubling trick can
1256 also be used to obtain anytime bounds that depend on D instead of D_{\max} (see e.g. [26]). This uses
1257 constant step-size OMD on time-horizons of doubling lengths until the unknown true T is reached.

1258 We can also obtain bounds that adapt to the sequence of observed subgradients of the form

$$R_T = \frac{D_{\max}^{1/r} r^{1/r} r_\star^{1/r_\star}}{\mu^{1/r}} \cdot \left(\sum_{i=1}^T \|g_i\|_\star^{r_\star} \right)^{1/r_\star}$$

1259 by using $\eta_t = D_{\max}^{1/r_\star} (r-1)^{1/r_\star} \mu^{1/r} / (\sum_{i=1}^T \|g_i\|_\star^{r_\star})^{1/r_\star}$. This follows the same lines as for OMD
1260 with strongly convex regulariser (see Section 4.2.1 of [35]).

1261 G.2 OMD on ℓ_p -balls

1262 • **For the low-dimensional setting**, consider OMD with regulariser $\phi_2(x) = \frac{1}{2}\|x\|_2^2$. We have

1263 $D_{\max} = \sup_{x,y \in \mathcal{B}_p} \frac{1}{2}\|x-y\|_2^2 = 2d^{1-2/p}$ and using that ϕ_2 is 1-strongly-convex with respect to

1264 $\|\cdot\|_2$, we have from (20) with $r = 2$ and $\eta_t = \frac{1}{L} \sqrt{\frac{2d^{1-2/p}}{t}}$:

$$R_T \leq 2L \sqrt{2d^{1-2/p} T}.$$

1265 • **For the high-dimensional setting**, consider OMD with regulariser $\phi_p(x) = \frac{1}{p}\|x\|_p^p$. We have

1266 $D_{\max} = \sup_{x,y \in \mathcal{B}_p} D_{\phi_p}(x,y) = 2$ (see below) and using that ϕ_p is 2^{1-p} -uniformly-convex of

1267 degree p with respect to $\|\cdot\|_p$, we have from (20) with $r = p$ and $\eta = \frac{1}{L} \left(\frac{p-1}{t} \right)^{1-1/p}$:

$$R_T \leq 2p^{1/p} p_\star^{1-1/p_\star} L T^{1-1/p}.$$

1268 To show $D_{\max} = 2$: Fix $x, y \in \mathcal{B}_p$ and $\psi(x) = \frac{1}{p}\|x\|_p^p$. The sign, power and absolute value
1269 functions below are applied component-wise to vectors.

$$\begin{aligned} D_\psi(x, y) &= \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle \\ &= \frac{1}{p}\|x\|_p^p - \frac{1}{p}\|y\|_p^p + \sum_{i=1}^d \left\{ \text{sign}(y_i) \cdot |y_i|^{p-1} (y_i - x_i) \right\} \\ &= \frac{1}{p}\|x\|_p^p - \frac{1}{p}\|y\|_p^p + \sum_{i=1}^d \left\{ |y_i|^p - \text{sign}(y_i) \cdot |y_i|^{p-1} x_i \right\} \\ &= \frac{1}{p}\|x\|_p^p + \left(1 - \frac{1}{p}\right)\|y\|_p^p - \sum_{i=1}^d \left\{ \text{sign}(y_i) \cdot |y_i|^{p-1} x_i \right\} \\ &\leq \frac{1}{p} + \left(1 - \frac{1}{p}\right) - \sum_{i=1}^d \left\{ \text{sign}(y_i) \cdot |y_i|^{p-1} x_i \right\}. \end{aligned}$$

1270 We show that the last term is bounded by 1 by using Holder's inequality,

$$\langle \text{sign}(y) \cdot |y|^{p-1}, x \rangle \leq \| |y|^{p-1} \|_q \|x\|_p \leq 1,$$

1271 where in the last inequality we used (recall that $q = p/(p-1)$)

$$\| |y|^{p-1} \|_q = \left(\sum_{i=1}^d |y_i|^{q \cdot (p-1)} \right)^{1/q} = \left(\sum_{i=1}^d |y_i|^p \right)^{1/q} = \|y\|_p^{p/q} \leq 1.$$

1272 Hence $\sup_{x,y \in \mathcal{B}_p} D_\psi(x, y) \leq 2 = D_{\max}$.

1273 • **We now show how to achieve anytime optimal bounds.** Fix $t_0 = \left(\sqrt{2} p^{1/p} p_\star^{1/p_\star} \right)^{2p/(p-2)} \cdot d$.

1274 **Proposition G.3.** Consider running OMD with the following regularizers

$$\psi_t(x) = \begin{cases} \phi_p(x) = \frac{1}{p}\|x\|_p^p, & \eta_t = \frac{(p-1)^{1/p_\star}}{L t^{1/p_\star}}, \quad \text{if } t \leq t_0, \\ \phi_2(x) = \frac{1}{2}\|x\|_2^2, & \eta_t = \frac{\sqrt{2d^{1-2/p}}}{L \sqrt{t}}, \quad \text{if } t > t_0. \end{cases}$$

1275 Assume ℓ_t convex, closed, and $\partial \ell_t(x_t)$ not empty. Then, OMD guarantees

$$R_T \leq \begin{cases} 2p^{1/p} p_\star^{1/p_\star} L T^{1/p_\star}, & \text{if } t \leq t_0, \\ 2L \sqrt{2T d^{1-2/p}}, & \text{if } t > t_0 \end{cases}$$

1276 *Proof.* If $T \leq t_0$, we have just run OMD with ϕ_p as regulariser over all rounds and the regret
 1277 bound is the one for the high-dimensional setting above.

1278 Otherwise, from the standard bounds from the OMD analysis

$$\begin{aligned}
 R_T &\leq \sum_{t=1}^{t_0} \left(\frac{D_{\psi_p}(u, x_t)}{\eta_t} - \frac{D_{\psi_p}(u, x_{t+1})}{\eta_t} \right) + \frac{L^{p^*}}{p_* \mu_p^{p^*-1}} \sum_{t=1}^T \eta_t^{p^*-1} \\
 &\quad + \sum_{t=t_0+1}^T \left(\frac{D_{\psi_2}(u, x_t)}{\eta_t} - \frac{D_{\psi_2}(u, x_{t+1})}{\eta_t} \right) + \frac{L^2}{2\mu_2} \sum_{t=t_0+1}^T \eta_t \\
 &\leq \frac{D_p}{\eta_{t_0}} + \frac{L^{p^*}}{p_* \mu_p^{p^*-1}} \sum_{t=1}^T \eta_t^{p^*-1} + \frac{D_2}{\eta_T} + \frac{L^2}{2\mu_2} \sum_{t=t_0+1}^T \eta_t \\
 &\leq 2p^{1/p} p_*^{1/p^*} L t_0^{1/p^*} + \frac{D_2}{\eta_T} + \frac{L\sqrt{D_2}}{2\mu} \sum_{t=t_0+1}^T \frac{1}{\sqrt{t}} \\
 &\leq 2p^{1/p} p_*^{1/p^*} L t_0^{1/p^*} + L\sqrt{TD_2} + L\sqrt{D_2}(\sqrt{T} - \sqrt{t_0}) \\
 &= 2L\sqrt{2Td^{1-2/p}} + 2p^{1/p} p_*^{1/p^*} L t_0^{1/p^*} - L\sqrt{2d^{1-2/p}t_0},
 \end{aligned}$$

1279 where we used $D_2 = \sup_{x,y \in \mathcal{B}_p} D_{\psi_2}(x, y) \leq 2d^{1-2/p}$, $D_p = \sup_{x,y \in \mathcal{B}_p} D_{\psi_p}(x, y) \leq 2$, $\mu_2 = 1$
 1280 and $\mu_p = 2^{1-p}$. The proof is concluded by noting that $2p^{1/p} p_*^{1/p^*} L t_0^{1/p^*} - L\sqrt{2d^{1-2/p}t_0}$ is
 1281 negative for $t_0 = \left(\sqrt{2} p^{1/p} p_*^{1/p^*} \right)^{2p/(p-2)} \cdot d$. \square

1282 G.3 Failure of fixed separable regularisation for OMD

1283 **Proposition G.4.** *OMD with regulariser $\psi \in \Psi$ and any sequence of decreasing η_t cannot be*
 1284 *optimal across all dimensions. Specifically there are no constants $c_h, c_l > 0$ such that for all T ,*
 1285 *$R_T \leq c_h L T^{1-1/p}$ for all $d > T$ and $R_T \leq c_l L \sqrt{T} d^{1-2/p}$ for all $d \leq T$.*

1286 The proof is identical to the proof of Theorem 4.6 for FTRL with the corresponding versions of
 1287 Lemma 4.4 and Lemma 4.7 for OMD given below.

1288 **Lemma G.5.** *[Lemma 4.7 for OMD] Consider $d = 1$ ($V = \mathcal{B}_p = [-1, 1]$) and $\psi \in \mathcal{F}$. OMD*
 1289 *with regulariser ψ and arbitrary decreasing step-size η_t can only guarantee $R_T \leq cL\sqrt{T}$ for some*
 1290 *constant $c > 0$ and all sufficiently large T if for all $x \in [-1, 1]$, $\psi(x) \geq \frac{\psi(1/2)}{100c^2} x^2$.*

1291 *Proof.* Assume there exists a constant $c > 0$ such that for all T and any sequence of losses, $R_T \leq$
 1292 $cL\sqrt{T}$. Consider $T > 16c^2$ and a multiple of 4.

1293 By considering the dual-version of OMD, we have that if there are no projections up to time t , the
 1294 update of OMD at time $t+1$ can be written as

$$x_{t+1} = \nabla \psi_V^* \left(- \sum_{i=1}^t \eta_i g_i \right) = \arg \min_{x \in V} \left\{ \psi(x) + \sum_{i=1}^t \eta_i g_i^T x \right\}, \quad (21)$$

1295 where ψ_V^* is the restriction of the fenchel conjugate of ψ to $V = \mathcal{B}_p = [-1, 1]$.

1296 We now follow the same steps as FTRL with a slight modification to the loss $\ell_t(x) = x \cdot g_t$ where
 1297 $g_t \in [-1, 1]$ is now defined as

$$g_t = \begin{cases} -\frac{\eta_{t+1}}{\eta_t} \cdot L, & t \leq \frac{T}{2} \text{ odd,} \\ L, & t \leq \frac{T}{2} \text{ even,} \\ -L, & t > \frac{T}{2}, \end{cases}$$

1298 Assume η_2 is small enough s.t. $x_2 \in \text{int } V = (-1, 1)$ (i.e. no projection is needed). If not, we can
 1299 modify the losses slightly so that $x_3 = 0$ (if η_3 is large enough, if it is not then set $g_1 = g_2 = 0$ and
 1300 start the above losses from $t = 3$) and then proceed similarly (i.e. if η_3 is still so large that $x_4 = 1$,
 1301 then again modify the losses slightly so that $x_5 = 0$ etc). Set $\eta = \eta_{T/2}$. With this sequence of losses,
 1302 the points played by OMD satisfy

- 1303 • for $t \leq T/2 + 1$ and t odd, we have $\sum_{s=1}^{t-1} \eta_s g_s = 0$, so $x_t = 0$.
- 1304 • for $t \leq T/2 + 1$ and t even, we have $\sum_{s=1}^{t-1} \eta_s g_s = -\eta_t \cdot L$ so $x_t = \arg \min_{x \in [-1, 1]} \{-\eta_t x + \psi(x)\}$.
- 1305 For $t < t' \leq T/2$ (both even), we have

$$\begin{aligned}
-\eta_{t'} L x_{t'} + \psi(x_{t'}) &\leq -\eta_{t'} L x_t + \psi(x_t) && \text{using the definition of } x_{t'} \\
&= -\eta_t L x_t + \psi(x_t) + L(\eta_t - \eta_{t'}) x_t \\
&\leq -\eta_t L x_{t'} + \psi(x_{t'}) + L(\eta_t - \eta_{t'}) x_t && \text{using the definition of } x_t \\
\implies (\eta_t - \eta_{t'}) L x_{t'} &\leq (\eta_t - \eta_{t'}) L x_t \\
\implies x_{t'} &\leq x_t && \text{using that } \eta_{t'} \leq \eta_t.
\end{aligned}$$

1306 So for all $t \leq T/2$ even, we have $x_t \geq x_{T/2}$.

- 1307 • for $t > T/2$, we have $\sum_{s=1}^{t-1} \eta_s g_s \geq \sum_{s=1}^t \eta_s g_s$ so $x_t \leq x_{t+1}$.

1308 The regret can be written as follows

$$R_T = \sum_{t=1}^T \ell_t(x_t) - \left(-\frac{LT}{2}\right) \geq \frac{LT}{2} + \frac{LT}{4} x_{T/2} - L \sum_{t=T/2+1}^T x_t \quad (22)$$

1309 Following similar steps as the proof of Lemma 4.7 for FTRL, we get

- 1310 1. We first show that $x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil} \geq \frac{1}{2}$: if not, $x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil} < \frac{1}{2}$ and from (22):

$$\begin{aligned}
R_T &\geq \frac{LT}{2} - L \sum_{t=T/2+1}^{T/2 + \lceil 2c\sqrt{T} \rceil} x_t - L \sum_{t=T/2 + \lceil 2c\sqrt{T} \rceil}^T x_t \\
&> \frac{LT}{2} - L \sum_{t=T/2+1}^{T/2 + \lceil 2c\sqrt{T} \rceil} \frac{1}{2} - L \sum_{t=T/2 + \lceil 2c\sqrt{T} \rceil + 1}^T 1 \\
&= \frac{LT}{2} - \frac{L}{2} \lceil 2c\sqrt{T} \rceil - L \left(T - \lceil 2c\sqrt{T} \rceil - \frac{T}{2} \right) \\
&\geq \frac{L}{2} \lceil 2c\sqrt{T} \rceil \\
&> cL\sqrt{T},
\end{aligned}$$

1311 which contradicts our initial assumption that $R_T \leq cL\sqrt{T}$ so we must have $x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil} \geq$

1312 $\frac{1}{2}$. Note that $2c\sqrt{T} < T/2$ is ensured by $T > 16c^2$.

- 1313 2. Next, we show that $\eta \geq \frac{\psi(1/2)}{2cL\sqrt{T}}$. Until the points reach 1 there are no projections so we can
- 1314 use (21) to write the OMD update as (note that even if $x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil} = 1$ the following still
- 1315 holds)

$$\begin{aligned}
x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil} &= \arg \min_{x \in V} \left\{ \psi(x) + \sum_{i=1}^{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil - 1} \eta_i g_i \cdot x \right\} = \arg \min_{x \in V} \left\{ \psi(x) - x \sum_{i=\frac{T}{2}+1}^{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil - 1} \eta_i \right\} \\
\implies -x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil} \sum_{i=\frac{T}{2}+1}^{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil - 1} \eta_i + \psi(x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil}) &\leq 0 \\
\implies \psi(1/2) &\leq \sum_{i=\frac{T}{2}+1}^{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil - 1} \eta_i \leq 2c\sqrt{T} \eta \\
\implies \eta &\geq \frac{\psi(1/2)}{2c\sqrt{T}}.
\end{aligned}$$

1316 where in the second implication, we used that $\psi(x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil}) \geq \psi(1/2)$ (since ψ is in-
 1317 creasing on $[0, 1]$ and $x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil} \geq 1/2$), $x_{\frac{T}{2} + \lceil 2c\sqrt{T} \rceil} \leq 1$ and $\eta_i \leq \eta_{T/2} = \eta$ for all
 1318 $i \geq T/2$.

1319 The remaining steps are identical to the proof of Lemma 4.7. \square

1320 **Lemma G.6** (Lemma 4.4 for OMD). *Consider $V = \mathcal{B}_p$ with $p > 2$ and assume losses are L -Lipschitz
 1321 in ℓ_p -norm. Let ψ be a convex function satisfying for some $\mu > 0$ and any $x \in \mathbb{R}^d$, $\psi(x) \geq \frac{\mu}{2} \|x\|_2^2$.
 1322 If $d \geq (4T/\mu)^{p/(p-2)}$, there exists a sequence of linear L -Lipschitz losses (in ℓ_p -norm) for which
 1323 OMD with regulariser $\psi(x)$ and any sequence of decreasing η_t suffers regret $R_T \geq \frac{1}{8}LT$.*

1324 *Proof.* We consider the loss construction described in Appendix E.1 with a slight modification to the
 1325 loss $\ell_t(x) = L \cdot x^T g_t$ where $g_t \in \mathcal{B}_{p^*}$ is now defined as

$$g_t = \begin{cases} -\frac{\eta_{t+1}}{\eta_t} L \cdot e_1, & t \leq \frac{T}{2} \text{ odd}, \\ L \cdot e_1, & t \leq \frac{T}{2} \text{ even}, \\ -L \cdot v, & t > \frac{T}{2}. \end{cases}$$

1326 By again considering the dual-version of OMD, we have that if there are no projections up to time t ,
 1327 the update of OMD at time $t + 1$ can be written as

$$x_{t+1} = \nabla \psi_V^* \left(-L \sum_{i=1}^t \eta_i g_i \right) = \arg \min_{x \in V} \left\{ \psi(x) + L \sum_{i=1}^t \eta_i g_i^T x \right\}, \quad (23)$$

1328 where ψ_V^* is the restriction of the fenchel conjugate of ψ to $V = \mathcal{B}_p$.

1329 • First we consider $t \leq T/2$. As in the proof of Lemma G.5, assume η_2 is small enough s.t.
 1330 $x_2 \in \text{int } \mathcal{B}_p$ (i.e. no projection is needed). By (23), the steps are then the same as for FTRL
 1331 since when t is odd, $x_t = 0$ and when t is even, $x_t = \arg \min_{x \in \mathcal{B}_p} \left\{ \psi(x) - L\eta_t e_1^T x \right\}$. So
 1332 we have

$$\sum_{t=1}^{T/2} x_t^T g_t \geq \begin{cases} T/4, & \text{if } \eta_{T/2} \geq 2/L \\ 0, & \text{if } \eta_{T/2} < 2/L. \end{cases}$$

1333 Hence if $\eta_{T/2-1} \geq 2/L$, we have $R_T \geq LT/4$ and the statement of the theorem holds. If
 1334 $\eta_{T/2-1} < 2/L$, we look to the second half of the rounds.

1335 • Let's now consider $t > T/2$ and assume $\eta_{T/2-1} < 2/L$. Fix $\beta_t = t - T/2 - 1$. Until the
 1336 points reaches the boundary there are no projections so we can use (21) to write the OMD
 1337 update as

$$\begin{aligned} x_t &= \arg \min_{x \in \mathcal{B}_p} \left\{ \psi(x) + L \sum_{i=1}^{t-1} \eta_i g_i^T x \right\} = \arg \min_{x \in \mathcal{B}_p} \left\{ \psi(x) - L \cdot v^T x \sum_{i=T/2+1}^{t-1} \eta_i \right\} \\ &= \arg \min_{x \in \mathcal{B}_p} \sum_{i=1}^d \left\{ g(x_i) - L x_i d^{-1/q} \sum_{i=T/2+1}^{t-1} \eta_i \right\}. \end{aligned}$$

1338 Let $u = v/\|v\|_p$ be the competitor. Note that $x_t = \lambda_t u$ (since the update is coordinate
 1339 invariant) so only reaches the boundary once $x_t = u$ and for which the above equality
 1340 still holds (this is true because it is true for the last iterate before the projection and then
 1341 $\sum_{i=1}^{t-1} \eta_i g_i$ is greater than for this last iterate so the argmin will give $x_t = u$). We have
 1342 $\lambda_t \geq 0$ and

$$\psi(x_t) \geq \frac{\mu}{2} \|x_t\|_2^2 = \frac{1}{2} \lambda_t^2 \mu d^{1-2/p}.$$

1343

Now from the OMD update,

$$\begin{aligned}
\psi(x_t) - Lv^T x_t \sum_{i=T/2+1}^{t-1} \eta_i \leq 0 &\implies \frac{1}{2} \lambda_t^2 \mu d^{1-2/p} - \lambda_t L \sum_{i=T/2+1}^{t-1} \eta_i \leq 0 \\
&\implies \lambda_t \leq \frac{2L}{\mu d^{1-2/p}} \sum_{i=T/2+1}^{t-1} \eta_i \leq \frac{2L\beta_t \eta}{\mu d^{1-2/p}} \leq \frac{4\beta_t}{\mu d^{1-2/p}} \\
&\implies \ell_t(x_t) = -Lv^T x_t = -L\lambda_t \geq -\frac{4L\beta_t}{\mu d^{1-2/p}},
\end{aligned}$$

1344

since $\eta_i \leq \eta_{T/2} \leq 2/L$ for all $i \geq T/2$. If $d \geq (4T/\mu)^{p/(p-2)}$, we have for all $t \leq T$

$$\begin{aligned}
\ell_t(x_t) &\geq -\frac{4L\beta_t}{\mu d^{1-2/p}} \geq -\frac{L}{2} \\
\implies R_T &\geq \frac{LT}{2} + \sum_{t=T/2+1}^T \ell_t(x_t) \geq \frac{LT}{2} - \frac{LT}{4} = \frac{LT}{4}.
\end{aligned}$$

1345 If T is not divisible by 4 and we use $T-1$, $T-2$ or $T-3$, we have $R_T \geq \frac{L(T-3)}{4} \geq \frac{LT}{8}$ for $T \geq 6$,
1346 concluding the proof. \square