Journal of Machine Learning Research 23 (2022) 1-12 Submitted 11/20; Revised 9/22; Published 9/22

Expected Regret and Pseudo-Regret are Equivalent When the Optimal Arm is Unique

Daron Anderson ANDERSD3@TCD.IE
Department of Computer Science and Statistics,
Trinity College Dublin, Ireland

Douglas J. Leith DOUG.LEITH@TCD.IE
Department of Computer Science and Statistics,
Trinity College Dublin, Ireland

Editor: Aurelien Garivier

Abstract

In online linear optimisation with stochastic losses it is common to bound the pseudo-regret of an algorithm rather than the expected regret. This is attributed to the expected fluctuations for i.i.d sums making expected regret bounds better than $\Omega(\sqrt{T})$ impossible. In this paper we show that when there is a unique optimal action and the action set is a polytope the difference between pseudo-regret and expected regret is $o(1)$. This means that the existing upper bounds on pseudo-regret in the literature can immediately be extended to also upper bound the expected regret. Our results are independent of the algorithm used to select the actions and apply equally to the bandit and full-information settings.

Keywords: online optimisation, bandits, regret

1. Introduction

In online linear optimisation we choose an action x_t each turn from the domain $X \subset \mathbb{R}^d$. The adversary then chooses a loss vector $\ell_t \in \mathbb{R}^d$ and we suffer loss $\ell_t \cdot x_t$. In the full-information setting we observe ℓ_t as feedback while in the bandit setting we only observe $\ell_t \cdot x_t$. Then we proceed to turn $t+1$. Our aim is to select actions that minimise the *regret* $R_T = \sum_{t=1}^T \ell_t \cdot x_t - \sum_{t=1}^T \ell_t \cdot x_T^*$ with respect to the benchmark $x_T^* \in \operatorname{argmin}_{x \in X} \sum_{t=1}^T \ell_t \cdot x$.

The standard model of an easy opponent is when $\ell_1, \ell_2, \dots$ are independent and identically distributed random variables with each $\mathbb{E}[\ell_t] = \bar{\ell}$. In this case superior performance is expected, and we would like to minimise the expected regret $\mathbb{E}[R_T] = \mathbb{E}\big[\sum_{t=1}^T \ell_t \cdot x_t\big] - \mathbb{E}\big[\sum_{t=1}^T \ell_t \cdot x_T^*\big]$ where the expectation is over the loss vectors. Here x_T^* is a random variable that depends on the realisations of $\ell_1, \ell_2, \dots, \ell_T$.

Unfortunately, the expected regret can be $\Omega\big(\sqrt{T}\big)$ in general, which matches the worst-case bound for adversarial losses. To remedy this, many classical performance bounds instead use the *pseudo-regret* $\overline{R}_T = \mathbb{E}\big[\sum_{t=1}^T \ell_t \cdot x_t\big] - \mathbb{E}\big[\sum_{t=1}^T \ell_t \cdot y^*\big]$ with benchmark $y^* \in \operatorname{argmin}\{\bar{\ell} \cdot x : x \in X\}$ that minimises the expected loss rather than the actualised loss. For example early work by Dani et al. (2008) established an $O(\log(T)/\Delta)$ upper bound on the pseudo-regret with bandit feedback, where Δ is the difference between the smallest and second smallest component of $\bar{\ell}$. This bound was subsequently improved by Abbasi-yadkori et al. (2011) while Auer and Chiang (2016) shows that a player can simultaneously achieve $O(\log(T)/\Delta)$ pseudo-regret for stochastic losses and $O(\sqrt{T})$ pseudo-regret for adversarial losses. In the full information setting Anderson and Leith (2020); Mourtada and Gaïffas (2019) establish an $O(1/\Delta)$ upper bound on the pseudo-regret when the

loss vectors are stochastic, and Anderson and Leith (2021) prove an $O(\log N)$ bound for strongly convex domains.

Since the benchmark $y^* \in \operatorname{argmin}\{\overline{\ell} \cdot x : x \in X\}$ from the pseudo-regret depends only on the average loss $\overline{\ell}$ it does not depend on the specific realisation of $\ell_1, \ell_2, \ldots, \ell_T$. This simplifies the analysis but comes at the price that $\overline{R}_T \leq \mathbb{E}[R_T]$. Hence an upper bound on pseudo-regret $\overline{R}_T$ need not imply an upper bound on expected regret $\mathbb{E}[R_T]$. Indeed, it is often claimed in the bandit literature that the difference between expected regret and pseudo-regret is always $\Omega(\sqrt{T})$. In this paper we show that in the favourable case of a unique optimal action and action set X a polytope this difference vanishes, i.e. $0 \leq \mathbb{E}[R_T] - \overline{R}_T \leq o(1)$. This means that the existing upper bounds on pseudo-regret in the literature can immediately be extended to also upper bound the expected regret. Our results are independent of the algorithm used to select the actions and apply equally to the bandit and full-information settings.

Polytope domains are used for example in ranking user preferences, travelling salesman problems, and assigning "vehicles" to "routes" in a transportation problem. In the more general setting of combinatorial optimisation, we seek an optimal action from a large finite set. For example we want to select a permutation of n elements. The nature of the loss function allows us to embed the actions as points in a lower-dimensional space. The polytope arises as the convex hull of the embedded actions, with each interior point treated as a probabilistic choice of the vertices. Then we can optimise over the polytope to obviate the computationally infeasible task of checking each of the $n!$ permutations individually. For examples of learning permutations see Zhang (2004); Helmbold and Warmuth (2009); Suehiro et al. (2012); Lim and Wright (2014); Ailon (2014); Ailon et al. (2016); Linderman et al. (2018) and the references therein. For other polytopes see Warmuth and Kuzmin (2008) and Kalai and Vempala (2016). For overviews see Martin Grötschel, László Lovász, Alexander Schrijver (1993) or Mark Hickman, Pitu Mirchandani, Stefan Voß (2008).

1.1 Previous Bounds on Expected Regret

Most of the stochastic online linear optimisation literature focuses on the pseudo-regret. However there are some notable exceptions. Gaillard et al. (2014) consider a new variant of the Prod algorithm on the simplex, with different learning rates for each arm and full information feedback. Their Theorem 11 shows that for i.i.d loss vectors the algorithm gives $\mathbb{E}[R_T] \leq O(1)$ provided the expected loss vector has a unique optimiser. Corollary 11 of Huang et al. (2016) says that the Follow-The-Leader algorithm with full information feedback gives $\mathbb{E}[R_T] \leq O(1)$ when the loss vectors are i.i.d, the domain is a polytope and the expected loss vector has a unique optimiser. For strongly convex domains their Theorem 5 implies that for i.i.d loss vectors running Follow-The-Leader gives $\mathbb{E}[R_T] \leq O(\log T)$. Wei and Luo (2018) consider problems on the simplex under bandit feedback. In Section 4.2 they give an algorithm with the standard $O(\sqrt{T \log T})$ regret bound against adaptive adversaries, which specialises to give $\mathbb{E}[R_T] \leq O(\log T/\Delta)$ when the loss vectors satisfy the martingale [1] inequality $\mathbb{E}[\ell_t(j) - \ell_t(j^*) \,|\, \ell_1, \ldots, \ell_{t-1}] > \Delta$ for fixed j^* and all $j \neq j^*$. For strongly-convex loss functions it is well-known we can get $O(\log T)$ regret by running online gradient descent (Hazan et al., 2007) or follow-the-leader (Cesa-Bianchi and Lugosi (2006) Section 3.2). If the functions are generated by an i.i.d sequence we immediately have $\mathbb{E}[R_N] \leq O(\log T)$.

The existing literature therefore creates the confusing situation where, on the one hand most papers say that obtaining bounds on the expected regret is impossible, while on the other hand a few papers do manage to bound the expected regret. However they do this without commenting

1. Martingales are slightly more general than i.i.d sequences. Many of the common concentration inequalities used for i.i.d sequences generalise immediately to martingales.

on whether this is surprising or not and without dwelling on the assumptions. The present paper is a first step towards clarifying the situation.

2. Notation

For vectors $x \in \mathbb{R}^d$ we write $x(j)$ for the j-component and $\|x\|$ for the Euclidean norm. By a *polytope* we mean the convex hull of a finite set. Polytopes are always compact and convex. For every polytope $\mathcal{P}$ there exists a unique finite set $\mathcal{V}$ of *vertices* such that $\mathcal{P}$ is the convex hull of $\mathcal{V}$ and not the convex hull of any proper subset of $\mathcal{V}$. -

Assumption 1 The domain $X \subset \mathbb{R}^d$ is compact and convex. The diameter is bounded by $D = \sup\{\|x - y\|, x, y \in X\}$.

Assumption 2 The loss vectors $\ell_1, \ell_2, \ldots \in \mathbb{R}^d$ are independent random variables with each $\mathbb{E}[\ell_t] = \overline{\ell}$. The expected optimiser $y^* = \operatorname{argmin}\{\overline{\ell} \cdot x : x \in X\}$ is unique. There are constants L, B such that each $\|\ell_t\| \leq L$ and $\|\ell_t - \overline{\ell}\| \leq B$ almost surely.

Note Assumption 1 holds if X is a polytope. Note also we can always take $B = 2L$ in Assumption 2. The expected regret of a sequence of losses and actions is defined as

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^T \ell_t \cdot x_t - \sum_{t=1}^T \ell_t \cdot x_T^*\right] \qquad x_T^* \in \operatorname*{argmin}_{x \in X} \sum_{t=1}^T \ell_t \cdot x_t.$$

The pseudo-regret is defined as

$$\overline{R}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_t \cdot x_t - \sum_{t=1}^T \ell_t \cdot y^*\right] \qquad y^* = \operatorname*{argmin}_{x \in X} \overline{\ell} \cdot x.$$

By definition of x_T^* we always have $\overline{R}_T \leq \mathbb{E}[R_T]$.

3. Preliminaries

Our first observation is that even though $\mathbb{E}[R_T]$ and $\overline{R}_T$ are defined in terms of on the action sequence, the gap $\mathbb{E}[R_T] - \overline{R}_T$ depends only on the loss vectors. This is because the first sums in the two definitions cancel with each other.

Lemma 1 We have $\mathbb{E}[R_T] - \overline{R}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_t \cdot (y^* - x_T^*)\right]$.

Hence the gap $\mathbb{E}[R_T] - \overline{R}_T$ is independent of our choice of online algorithm. Indeed it is independent of whether we are in the full-information or bandit setting. Next we use Lemma 1 to bound the gap in terms of the probability that the realised optimiser coincides with the expected optimiser:

Lemma 2 For $D = \sup\{\|x - y\| : x, y \in X\}$ we have the bounds

$$\mathbb{E}[R_T] - \overline{R}_T \leq \sqrt{\mathbb{E}\|x_T^* - y^*\|^2}\sqrt{\mathbb{E}\left\|\sum_{t=1}^T (\ell_t - \bar{\ell})\right\|^2} \leq D\sqrt{P(x_T^* \neq y^*)}\sqrt{\mathbb{E}\left\|\sum_{t=1}^T (\ell_t - \bar{\ell})\right\|^2}$$

Proof Expand the right-hand-side of Lemma 1 to get

$$\mathbb{E}[R_T] - \overline{R}_T = \mathbb{E}\left[\sum_{t=1}^T \bar{\ell} \cdot (y^* - x_T^*)\right] + \mathbb{E}\left[\sum_{t=1}^T (\ell_t - \bar{\ell}) \cdot (y^* - x_T^*)\right].$$

Since y^* minimises $\bar{\ell} \cdot x$ the first term is nonpositive and Cauchy-Schwarz gives

$$\mathbb{E}[R_T] - \overline{R}_T \leq \mathbb{E}\left[(y^* - x_T^*) \cdot \sum_{t=1}^{T} (\ell_t - \bar{\ell})\right] \leq \mathbb{E}\left[\|y^* - x_T^*\| \Big\| \sum_{t=1}^{T} (\ell_t - \bar{\ell})\Big\|\right] \tag{1}$$

Next use Cauchy-Schwarz with the L^2 inner product $(f, g) = \mathbb{E}_\omega[f(\omega)g(\omega)]$ to bound the right-hand-side and get

$$\mathbb{E}[R_T] - \overline{R}_T \leq \sqrt{\mathbb{E}\|y^* - x_T^*\|^2} \sqrt{\mathbb{E}\big\| \textstyle\sum_{t=1}^{T} (\ell_t - \bar{\ell})\big\|^2}.$$

Since $\|x_T^* - y^*\|^2 = 0$ for $x_T^* = y^*$ and $\|x_T^* - y^*\|^2 \leq D^2$ otherwise, the first factor on the right is at most $\sqrt{D^2 P(x_T^* \neq y^*)} = D\sqrt{P(x_T^* \neq y^*)}$. ■

To bound the second factor in Lemma 2 we use the following vector concentration inequality that is a special case of Pinelis (1994) Theorem 3.5.

Theorem 1 Suppose $X_1, X_2, \ldots \in \mathbb{R}^d$ are independent random variables with expectation zero and each $\|X_n\| \leq B$. For each $\varepsilon \geq 0$ we have

$$P\left(\Big\| \sum_{t=1}^{T} X_t \Big\| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{\varepsilon^2}{2TB^2}\right)$$

The full version of the Pinelis theorem applies to martingales that take values in so-called $(2, D)$-smooth Banach spaces. This allows us to generalise our main theorem. See Section 6 for discussion.

4. Polytopes

Henceforth $\mathcal{P} \subset \mathbb{R}^d$ is a polytope with vertex set $\mathcal{V}$. Since every linear function on a polytope is minimised on a vertex we know $y^* \in \mathcal{V}$ and write $v^* = y^*$ to stress this. We also assume x_T^* is selected to be a vertex. Define the suboptimality gap $\Delta = \min\left\{\bar{\ell} \cdot (v - v^*) : v \in \mathcal{V}/\{v^*\}\right\}$.

Theorem 2 Suppose we have an online optimisation problem on a polytope domain, and the loss vectors satisfy Assumption 2. For expected regret $\mathbb{E}[R_T]$ and pseudo-regret $\overline{R}_T$ we have:

$$\mathbb{E}[R_T] - \overline{R}_T \leq o(1) \qquad\qquad \mathbb{E}[R_T] - \overline{R}_T \leq (5/2)D^2B^2/\Delta.$$

Proof Write $L_T = \frac{1}{T}\sum_{t=1}^{T} \ell_t$ for the normalised cumulative loss vector. Then $\mathbb{E}[L_T] = \bar{\ell}$. Suppose we have $\|L_T - \bar{\ell}\| < \Delta/D$. Then for each $v \in \mathcal{V}/\{v^*\}$ we have

$$\begin{aligned} L_T \cdot (v - v^*) &= \bar{\ell} \cdot (v - v^*) + (L_T - \bar{\ell}) \cdot (v - v^*) > \Delta + (L_T - \bar{\ell}) \cdot (v - v^*) \\ &\geq \Delta - \|L_T - \bar{\ell}\| \|v - v^*\| > \Delta - \frac{\Delta}{D} D = 0 \end{aligned} \tag{2}$$

where the second line uses Cauchy-Schwarz. The above shows $L_T \cdot (v - v^*) > 0$. Hence $L_T \cdot x$ is uniquely minimised at v^* and so $x_T^* = v^*$.

Now let $X_t = \ell_t - \bar{\ell}$ and $\varepsilon = T\Delta/D$. Since $\frac{1}{T}\sum_{t=1}^{T} X_t = L_T - \bar{\ell}$ our Theorem 1 says the event $\|L_T - \bar{\ell}\| < \Delta/D$ fails with probability at most $2\exp(-\Delta^2 T/2B^2D^2)$. Hence $P(x_T^* \neq v^*) \leq 2\exp(-\Delta^2 T/2B^2D^2)$ and Lemma 2 gives

$$\mathbb{E}[R_T] - \overline{R}_T \leq D\sqrt{P(x_T^* \neq v^*)} \sqrt{\mathbb{E}\big\| \textstyle\sum_{t=1}^{T} (\ell_t - \bar{\ell})\big\|^2} \leq D\sqrt{2} \exp\left(-\frac{\Delta^2}{4B^2D^2} T\right) \sqrt{\mathbb{E}\big\| \textstyle\sum_{t=1}^{T} (\ell_t - \bar{\ell})\big\|^2}.$$

To bound the second factor use Theorem 1 to get $P\left(\left\|\sum_{t=1}^T(\bar{\ell}-\ell_t)\right\|>\varepsilon\right)\le 2\exp\left(-\varepsilon^2/2TB^2\right)$ for each $\varepsilon\ge 0$. For $x=\varepsilon^2$ we get $P\left(\left\|\sum_{t=1}^T(\bar{\ell}-\ell_t)\right\|^2>x\right)\le 2\exp\left(-x/2TB^2\right)$ and Lemma 6 gives

$$\mathbb{E}\left[\left\|\sum_{t=1}^T(\bar{\ell}-\ell_t)\right\|^2\right]\le\int_0^\infty P\left(\left\|\sum_{t=1}^T(\bar{\ell}-\ell_t)\right\|^2>x\right)dx\le 2\int_0^\infty\exp\left(-\frac{x}{2TB^2}\right)dx=4TB^2.$$

Hence we get

$$\mathbb{E}[R_T]-\overline{R}_T\le D\sqrt{2}\exp\left(-\frac{\Delta^2}{4B^2D^2}T\right)\sqrt{4TB^2}=2\sqrt{2}BD\sqrt{T}\exp\left(-\frac{\Delta^2}{4B^2D^2}T\right).$$

To prove the right-hand side is $o(1)$ recall that, since any exponential function grows faster than any polynomial, we have $\sqrt{T}e^{-\beta T}\to 0$ as $T\to\infty$ for any $\beta\ge 0$. To prove $\mathbb{E}[R_T]-\overline{R}_T\le(5/2)D^2B^2/\Delta$. Lemma 5 says $\sqrt{T}e^{-\beta T}\le 1/\sqrt{2e\beta}$. Hence for $\beta=\Delta^2/4B^2D^2$ the above gives

$$\mathbb{E}[R_T]-\overline{R}_T\le 2\sqrt{2}BD\sqrt{T}e^{-\beta T}\le\frac{2\sqrt{2}BD}{\sqrt{2e}}\sqrt{\frac{4B^2D^2}{\Delta^2}}=\frac{4}{\sqrt{e}}\frac{B^2D^2}{\Delta}\le\frac{5}{2}\frac{B^2D^2}{\Delta}.$$

■

5. Examples

In Theorem 2 we assume (a) the expected loss vector has a unique optimiser and (b) the domain is a polytope. Under these assumptions we prove the gap $\mathbb{E}[R_T]-\overline{R}_T$ tends to zero. In this section we give examples to show that both assumptions are needed. If we drop either of the two assumptions, we risk the gap becoming $\Omega(\sqrt{T})$.

5.1 A Unique Optimiser is Necessary

The following toy example demonstrates that if several optimisers are allowed we might have $\mathbb{E}[R_T]-\overline{R}_T\ge\Omega(\sqrt{T})$. i.e. the gap is the same size as the expected fluctuation $\mathbb{E}\left\|\sum_{t=1}^T(\ell_t-\bar{\ell})\right\|$ between the actual and expected loss vectors. This also matches the worst-case regret bound against adversarial losses.

Lemma 3 There exists an i.i.d sequence $\ell_1,\ell_2,\ldots\in\mathbb{R}$ of loss vectors with all $\|\ell_t\|,\|\ell_t-\bar{\ell}\|\le 1$ such that for domain the interval $[0,1]$ we have $\mathbb{E}[R_T]-\overline{R}_T\ge\sqrt{T}/10$ for T sufficiently large.

Proof Let ℓ_t be i.i.d with each $P(\ell_t=1)=P(\ell_t=-1)=1/2$. Then we have $\bar{\ell}=0$. Lemma 1 says $\mathbb{E}[R_T]-\overline{R}_T=\mathbb{E}\left[\sum_{t=1}^T\ell_t(y^*-x_T^*)\right]$ where y^* minimises $\bar{\ell}$ and x_T^* minimises $\sum_{t=1}^T\ell_t x$ over $[0,1]$. Consider the expression $S=\sum_{t=1}^T\ell_t(y^*-x_T^*)$ inside the expectation. Since $\bar{\ell}=0$ we can take $y^*=0$ to get $S=-\sum_{t=1}^T\ell_t x_T^*$. In the event that $\sum_{t=1}^T\ell_t<0$ we have $x_T^*=1$. In particular the event $\frac{1}{\sqrt{T}}\sum_{t=1}^T\ell_t<-1$ implies $S=-\sum_{t=1}^T\ell_t\ge\sqrt{T}$.

By the central limit theorem $\frac{1}{\sqrt{T}}\sum_{t=1}^T\ell_t$ converges in distribution to a standard normal. Hence as $T\to\infty$ we have $P\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T\ell_t<-1\right)\to P(N(0,1)<-1)$. The right-hand-side can be computed numerically as $0.158655...\ge 0.15$. Hence with probability at least 0.15 we have $S\ge\sqrt{T}$.

Since S is nonnegative we get $\mathbb{E}[R_T] - \overline{R}_T = \mathbb{E}[S] \geq 0.15\sqrt{T} \geq \sqrt{T}/10$. ■

Note that while the example in Lemma 3 seems trivial, it is in fact characteristic of all problems where Theorem 2 fails. This is because every problem with several optimisers contains a copy of the example. In all such problems the domain has vertices w, v such that the behaviour of the loss vectors over the line segment from w to v behaves like the interval $[0, 1]$ in the above example. We sketch the proof below.

Suppose $\mathcal{P} \subset \mathbb{R}^d$ is a polytope such that (a) the face $F^* = \{x \in \mathcal{P} : \overline{\ell} \cdot x = \overline{\ell} \cdot v^*\}$ of expected optimisers has dimension $m > 1$ and (b) the projections a_t of $\ell_t - \overline{\ell}$ onto the affine hull A of F^* are not essentially zero. By translating we can assume A is a hyperplane through the origin. Write $\{w, v_2, \ldots, v_M\}$ for the vertices of F^*. Since F^* has full dimension in A the vectors $w - v_j$ span A and we can extract a basis, without loss of generality $\mathcal{B} = \{w - v_j : j = 2, \ldots, m\}$. Then $\|x\|_{\mathcal{B}} = \max\{|x \cdot (w - v_j)| : j = 1, \ldots, m\}$ is a norm on A.

By assumption (b) there are $\varepsilon', p > 0$ such that $P(\|a_t\| > \varepsilon') > p$. Since all norms on $\mathbb{R}^m$ are equivalent we have $P(\|a_t\|_{\mathcal{B}} > \varepsilon) > p$ for some $\varepsilon > 0$. Hence with probability p we have some $|a_t \cdot (w - v_j)| > \varepsilon$. Since there are $m - 1$ choices for j there exists some fixed $v = v_j$ with $P(|a_t \cdot (w - v)| > \varepsilon) > p/m$. By symmetry we can assume $P(a_t \cdot (w - v) < -\varepsilon) > p/2m$.

Since $w - v \in A$ and a_t is the projection of ℓ_t onto the plane A we have $a_t \cdot (w - v) = \ell_t \cdot (w - v)$ and so $P(-\ell_t \cdot (w - v) > \varepsilon) > p/2m$. In particular $-\ell_t \cdot (w - v)$ are essentially nonzero random variables. Since $v, w \in F^*$ we have $\mathbb{E}[-\ell_t \cdot (w - v)] = 0$.

By the central limit theorem $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \ell_t \cdot (w - v)$ converges in distribution to some $N(0, \sigma)$ where $\sigma > 0$ depends on ε and $p/2m$. Hence for T sufficiently high we have $P\left(-\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \ell_t \cdot (w - v) > \sigma\right) > 0.15$. Like in Lemma 3 consider the expression $S = \sum_{t=1}^{T} \ell_t \cdot (y^* - x_T^*)$. We can take y^* as any element of F^*. For $y^* = v$ we have

$$S = \sum_{t=1}^{T} \ell_t \cdot v - \min_{x \in \mathcal{P}} \sum_{t=1}^{T} \ell_t \cdot x = \max_{x \in \mathcal{P}} \left(\sum_{t=1}^{T} \ell_t \cdot v - \sum_{t=1}^{T} \ell_t \cdot x \right)$$
$$= \max_{x \in \mathcal{P}} \left(-\sum_{t=1}^{T} \ell_t \cdot (x - v) \right) \geq -\sum_{t=1}^{T} \ell_t \cdot (w - v)$$

We have already shown the right-hand-side is greater than $\sigma\sqrt{T}$ with probability 0.15. Since S is nonnegative we get $\mathbb{E}[R_T] - \overline{R}_T = \mathbb{E}[S] \geq 0.15\sigma\sqrt{T} \geq \sigma\sqrt{T}/10$.

5.2 A Unique Minimizer Is Not Sufficient

In Theorem 2 we assume the expected optimiser is unique and the domain is a polytope. Lemma 3 shows the first assumption is necessary, by giving a polytope with several distinct optimisers for which the theorem fails. Lemma 4 tackles the second assumption. It shows that, even if the expected optimiser is unique, the theorem can fail if the domain is not a polytope.

Lemma 4 Let $\varepsilon > 0$ be arbitrary. There exists a compact convex domain in $\mathbb{R}^2$ and i.i.d sequence $\ell_1, \ell_2, \ldots \in \mathbb{R}^2$ of loss vectors with all $\|\ell_t\|, \|\ell_t - \overline{\ell}\| \leq \sqrt{2}$ such that $\mathbb{E}[R_T] - \overline{R}_T \geq \dfrac{T^{\frac{1}{2}-\varepsilon}}{20}$ for T sufficiently large.

Proof Let $\alpha \geq 3$ be such that $\frac{1}{2(\alpha-1)} < \varepsilon$. Let the domain be $X = X_1 \cap X_2$ for $X_1 = \{(x, y) \in \mathbb{R}^2 : y \geq |x|^\alpha\}$ and $X_2 = \{(x, y) \in \mathbb{R}^2 : y \leq 1\}$. Since $|x|^\alpha$ is a convex function the domain is convex. Since $X \subset [-1, 1] \times [0, 1]$ the domain is bounded hence compact. Let the loss vectors be $\ell_t = (B_t, 1)$

for independent B_t with $P(B_t=1) = P(B_t=-1) = 1/2$. Then $\overline{\ell} = (0,1)$ with expected optimiser $y^* = (0,0)$. Write $B(T) = B_1 + \ldots + B_T$ and $L_T = \sum_{t=1}^T \ell_t = (B(T), T)$ for the cumulative loss vector.

Lemma 1 says $\mathbb{E}[R_T] - \overline{R}_T = \mathbb{E}\left[L_T \cdot (y^* - x_T^*)\right]$ where x_T^* minimises $L_T \cdot x$. Since $y^* = 0$ it is enough to show $-\mathbb{E}\left[L_T \cdot x_T^*\right] \geq \Omega(\sqrt{T})$. To determine x_T^* is is enough by symmetry of the domain to consider the case $B(T) \leq 0$. We claim x_T^* is the point $(x_0, |x_0|^\alpha)$ on the graph for $x_0 = (-B(T)/\alpha T)^{1/(\alpha-1)}$.

Since $|x|^\alpha$ is strictly convex it is enough to show $|x| < 1$ and L_T is normal inwards at the point above. For any $x \in [0,1)$ the slope at the point $(x, |x|^\alpha)$ of the graph is $\alpha x^{\alpha-1}$ and the inward normal is along $(-\alpha x^{\alpha-1}, 1)$. Rescale to get $(-T\alpha|x|^{\alpha-1}, T)$. Plug in x_0 and use how $|B(T)| = -B(T)$ to get $(-T\alpha|x|^{\alpha-1}, T) = L_T$ as claimed. Hence we have

$$\begin{aligned}
L_T \cdot x_T^* &= \binom{B(T)}{T} \left(\left(\frac{-B(T)}{\alpha T} \right)^{\frac{1}{\alpha-1}}, \left(\frac{-B(T)}{\alpha T} \right)^{\frac{\alpha}{\alpha-1}} \right) \\
&= B(T) \left(\frac{-B(T)}{\alpha T} \right)^{\frac{1}{\alpha-1}} + T \left(\frac{-B(T)}{\alpha T} \right)^{\frac{\alpha}{\alpha-1}} \\
&= -\alpha^{\frac{1}{1-\alpha}} T^{\frac{1}{1-\alpha}} |B(T)|^{1+\frac{1}{\alpha-1}} + \alpha^{\frac{\alpha}{1-\alpha}} T^{1+\frac{\alpha}{1-\alpha}} |B(T)|^{\frac{\alpha}{\alpha-1}} \\
&= -\alpha^{1+\frac{\alpha}{1-\alpha}} T^{\frac{1}{1-\alpha}} |B(T)|^{\frac{\alpha}{\alpha-1}} + \alpha^{\frac{\alpha}{1-\alpha}} T^{\frac{1}{1-\alpha}} |B(T)|^{\frac{\alpha}{\alpha-1}} \\
&= (1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}} T^{\frac{1}{1-\alpha}} |B(T)|^{\frac{\alpha}{\alpha-1}} \qquad (3)
\end{aligned}$$

In Lemma 3 we showed $P\left(B(T) \leq -\sqrt{T}\right) > 1/10$ for T sufficiently high. When the event occurs we have $|B(T)|^{\frac{\alpha}{\alpha-1}} \geq T^{\frac{\alpha}{2(\alpha-1)}}$ and $T^{\frac{1}{1-\alpha}} |B(T)|^{\frac{\alpha}{\alpha-1}} \geq T^{\frac{1}{1-\alpha}} T^{\frac{\alpha}{2(\alpha-1)}} = T^{\frac{1}{2} - \frac{\alpha}{2(\alpha-1)}}$. Hence with probability at least $1/10$ we have

$$-L_T \cdot x_T^* = (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} T^{\frac{1}{1-\alpha}} |B(T)|^{\frac{\alpha}{\alpha-1}} \geq (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} T^{\frac{1}{2} - \frac{\alpha}{2(\alpha-1)}} \geq (\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} T^{\frac{1}{2}-\varepsilon}$$

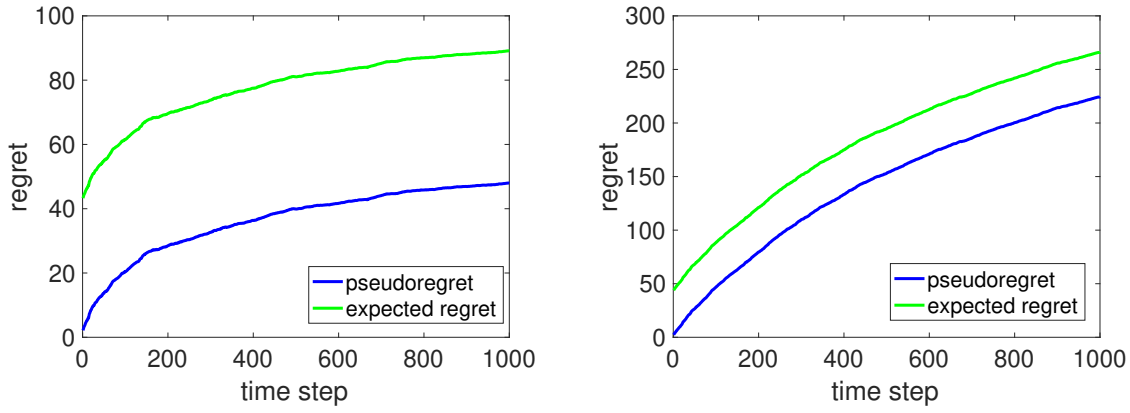
Since $-L_T \cdot x_T^*$ is nonnegative we get $\mathbb{E}[R_T] - \overline{R}_T \geq \frac{1}{10}(\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} T^{\frac{1}{2}-\varepsilon}$. To complete the proof we claim $(\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} \geq 1/2$ for α sufficiently high. To that end write $(\alpha-1)\alpha^{\frac{\alpha}{1-\alpha}} = \left(1 - \frac{1}{\alpha}\right) \alpha^{1+\frac{\alpha}{1-\alpha}} = \left(1 - \frac{1}{\alpha}\right) \alpha^{\frac{1}{1-\alpha}}$. The first factor tends to 1. To see the second factor tends to 1 observe $\log \alpha^{\frac{1}{1-\alpha}} = \frac{\log \alpha}{1-\alpha} = -\frac{\log \alpha}{\alpha-1}$. Since $\alpha - 1$ grows faster than $\log \alpha$ we have $\log \alpha^{\frac{1}{1-\alpha}} \to 0$ and so $\alpha^{\frac{1}{1-\alpha}} \to 1$. ■

5.3 Numerical Example

We present a brief example illustrating the application of our results. Similar to the discussion at the end of Section 5.1, this simple example captures all of the key behaviour.

In the usual multi-armed bandit setup the loss incurred by taking action x at time t is $\bar{\ell} \cdot x + \eta_t$ where $\eta_t \in \mathbb{R}$ is i.i.d and the action set X is a polytope with unit vectors as vertices. In other words, the loss when arm i is pulled at time t is $\bar{\ell}(i) + \eta_t$. Hence $\operatorname{argmin}_{x \in X} \left(\sum_{t=1}^T \bar{\ell} \cdot x + \eta_t \right) = \operatorname{argmin}_{x \in X} \sum_{t=1}^T \bar{\ell} \cdot x$ since the η_t term does not depend on x, and so the expected regret equals the pseudo-regret.

In the linear bandit setup we instead have loss $\ell_t \cdot x = (\bar{\ell} + \eta_t) \cdot x = \bar{\ell} \cdot x + \eta_t \cdot x$ where $\eta_t \in \mathbb{R}^d$ are i.i.d. vectors. The product with x in the random term gives $\operatorname{argmin}_{x \in X} \sum_{t=1}^T (\bar{\ell} + \eta_t) \cdot x \neq \operatorname{argmin}_{x \in X} \sum_{t=1}^T \bar{\ell} \cdot x$ in general. Hence the expected regret does not equal the pseudo-regret. Nevertheless, our analysis in this paper shows that the difference between the expected regret and pseudo-regret is uniformly



(a) Bandit Feedback

(b) Full Information Feedback

Figure 1: Example illustrating constant gap between expected regret and pseudo-regret.

bounded. We can therefore apply standard multi-arm bandit analysis techniques to upper bound pseudo-regret, and by extension upper bound the expected regret.

Figure 1 shows numerical simulations of a 2-arm linear bandit setup with $d = 2$, X the polytope with vertices $(1, 0)$, $(0, 1)$ and loss vector $\ell_t = (0, -24.9)$ with probability 0.5 and $(0, 25.1)$ with probability 0.5. Hence the expected loss is $\overline{\ell} = (0, 0.1)$ with optimiser $y^* = (1, 0)$. The pseudo-regret is $\mathbb{E}\big[\sum_{t=1}^T \ell_t \cdot x_t\big] - \overline{\ell} \cdot y^* = \mathbb{E}\big[\sum_{t=1}^T \ell_t \cdot x_t\big]$. The expected regret is $\mathbb{E}\big[\sum_{t=1}^T \ell_t \cdot (x_t - x_T^*)\big]$ where x_T^* minimises the sample path sum $\sum_{t=1}^T \ell_t \cdot x = \sum_{t=1}^T (\overline{\ell} + \eta_t) \cdot x$ and so varies from run to run. Figure 1(a) shows the measured pseudo-regret and expected regret for the UCB algorithm with bandit feedback. Figure 1(b) shows the corresponding values for the Lazy Gradient Descent algorithm of Anderson and Leith (2020) with full information feedback. It can be seen that, as expected, in both cases the difference between the expected regret and pseudo-regret is uniformly bounded and so the existing analytic results in terms of the pseudo-regret can be immediately generalised to encompass the expected regret.

6. Variants of the Main Theorem

Theorem 2 requires the domain to be a polytope. Our first observation is that, since we only use the polytope assumption in equation (2) of the proof, it is enough to assume the domain X and expected loss $\overline{\ell}$ admit a nonzero suboptimality gap. i.e there is a unique point $y^* \in X$ such that $\overline{\ell} \cdot (x - y^*) > \Delta > 0$ for all $x \in X / \{y^*\}$. For example the domain whose boundary is the union of the semicircular arc from $(1, 0)$ to $(-1, 0)$ and the line segments from $(-1, 0)$ to $(0, -1)$ to $(1, 0)$ satisfies the theorem if the loss vectors are such that $y^* = (0, -1)$.

The second set of observations comes from using the full version of Pinelis (1994) Theorem 3.5 rather than our special case of Theorem 1. This allows us to replace the independent losses in Theorem 2 with more general martingale losses. It also allows us to replace the Euclidean norm with any $(2, D)$-smooth norm. These are norms that satisfy a weakening of the parallelogram law. For a more general norm $|\cdot|$ we can replace the use of Cauchy-Schwarz in (2) with the fact $(L_T - \overline{\ell}) \cdot (v - v^*) \geq -|L_T - \overline{\ell}||v - v^*|^*$ for $|\cdot|^*$ the dual norm. This lets us prove a variant of Theorem 2 with the quantities D, B in Assumptions 1 and 2 replaced with $\sup |\ell_t - \overline{\ell}|$ and $\sup\{|x - y|^* : x, y \in X\}$ respectively. For example Pinelis Proposition 2.1 says the $\|\cdot\|_p$ norm is $(2, \sqrt{p-1})$ smooth. It is well known the dual norm is the Hölder conjugate.

For some norms that are not $(2, D)$ smooth it is still possible to prove analogous bounds to Theorem 1. For example suppose the domain is the simplex and we have $\|\ell_t - \overline{\ell}\|_\infty \leq B_\infty$. We can apply Azuma-Hoeffding to each component separately to see L_T is uniquely minimised at v^* except on a set with probability at most $\max\left\{1, d\exp\left(-\frac{\Delta^2}{8B_\infty^2}T\right)\right\}$. Using Lemma 6 to compute the expectation by integrating separately over two suitably chosen intervals will then yield a version of Theorem 2 with the quantities D, B replaced with 2 and B_∞ respectively.

Finally we remark that replacing the finite quantities D, B with finite quantities specialised to other norms is less significant than showing $\mathbb{E}[R_T] - \overline{R}_T \leq o(1)$ in the first place. This is because all norms are equivalent on a finite-dimensional vector space. Hence if $\sup|\ell_t - \overline{\ell}|$ and $\sup\{|x - y|^* : x, y \in X\}$ exist for any norm $|\cdot|$ then the Euclidean quantities B, D also exist and are finite. Hence we can apply Theorem 2 to see $\mathbb{E}[R_T] - \overline{R}_T \to 0$ without going to the trouble to reprove the theorem for the new norm.

7. Summary and Conclusions

In this paper we show that when there is a unique optimal action and the action set X is a polytope the difference between pseudo-regret and expected regret is $o(1)$. This means that the existing upper bounds on pseudo-regret in the literature can immediately be extended to also upper bound the expected regret. Our results are independent of the algorithm used to select the actions and apply equally to the bandit and full-information settings. This analysis can be extended to include i.i.d convex loss functions rather than just linear loss functions.

Importantly, while uniqueness of the optimal action is necessary for pseudo-regret and expected regret to be within $o(1)$ of one another, it is not sufficient and additional assumptions are needed. Here we use the additional assumption that the action set is a polytope. Another compatible assumption is that the action set is smooth and strongly convex. See Anderson and Leith (2021) Section 4. Other choices for the action set are likely possible.

Another direction for future work is to make stronger assumptions on the loss vectors, to allow a wider range of domains. For example the Bernstein condition (see Bartlett and Mendelson (2006) and van Erven et al. (2015)) quantifies how well the losses of almost optimal actions are correlated with the loses of optimal actions. For works about online problems under the Bernstein condition see Koolen and Van Erven (2015); van Erven et al. (2015); Koolen et al. (2016); van Erven and Koolen (2016); Mourtada and Gaïffas (2019).

Acknowledgments

This work was supported by SFI grant 16/IA/4610(T). We would also like to express out thanks for the helpful comments from the anonymous reviewers.

Appendix

Lemma 5 Let $A \geq 0$. The function $F(X) = \sqrt{X}e^{-AX}$ is maximised at $X = 1/2A$ and $F(1/2A) = 1/\sqrt{2eA}$. The function is increasing on $[0, 1/2A]$ and decreasing on $[1/2A, \infty)$.

Proof Consider the function $G(X) = Xe^{-2AX}$. The derivative $G'(X) = (1 - 2AX)e^{-2AX}$ is positive for before $X \leq 1/2A$ and vanishes for $X = 1/2A$ and is negative for $X \geq 1/2A$. Hence the function is increasing on $[0, 1/2A]$ and decreasing on $[1/2A, \infty)$ and maximised at $X = 1/2A$. Since the square root is monotone, the same holds for $\sqrt{G(X)} = F(X)$. The maximum value is

$F(1/2A) = \sqrt{1/2A}e^{-A(1/2A)} = e^{-A(1/2A)}/\sqrt{2A} = e^{-1/2}/\sqrt{2A} = 1/\sqrt{2eA}$ ■

The following fact about computing the expectation in terms of the CDF is well-known. But we were unable to find a suitably general proof in the literature.

Lemma 6 Suppose X is a nonnegative random variable. Then $\mathbb{E}[X] = \int_0^\infty P(X > x)dx$.

Proof The integral can be written as

$$\int_0^\infty P(X > x)dx = \int_0^\infty \mathbb{E}_y\big[\mathbf{1}_{X(y)>x}(y)\big]\, dx = \mathbb{E}_y\bigg[\int_0^\infty \mathbf{1}_{X(y)>x}(y)dx\bigg].$$

For fixed y define the function $g(x) = \mathbf{1}_{X(y)>x}(y)$. We have $g(x) = 1$ for all $x > X(y)$ and $g(x) = 0$ elsewhere. Since $X(y)$ is nonnegative that means $g(x)$ is the indicator function of $[0, X(y))$. It follows the inner integral equals $X(y)$ and the above becomes $\mathbb{E}_y[X(y)] = \mathbb{E}[X]$. ■

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

Nir Ailon. Improved Bounds for Online Learning Over the Permutahedron and Other Ranking Polytopes. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 29–37, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL https://proceedings.mlr.press/v33/ailon14.html.

Nir Ailon, Kohei Hatano, and Eiji Takimoto. Bandit online optimization over the permutahedron. *Theoretical Computer Science*, 650:92–108, 2016. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2016.07.033. URL https://www.sciencedirect.com/science/article/pii/S0304397516303784. Algorithmic Learning Theory.

Daron Anderson and Douglas Leith. Online Lazy Gradient Descent is Universal on Strongly Convex Domains. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5874–5884. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/2e907f44e0a9616314cf3d964d4e3c93-Paper.pdf.

Daron Anderson and Douglas J. Leith. Lazy Online Gradient Descent is Universal on Polytopes. *CoRR*, abs/2004.01739, 2020. URL https://arxiv.org/abs/2004.01739.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Proc Conference on Learning Theory (COLT)*, pages 116–120, 2016.

Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, July 2006. ISSN 1432-2064. doi: 10.1007/s00440-005-0462-3. URL https://doi.org/10.1007/s00440-005-0462-3.

Nicolò Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Proc Conference on Learning Theory (COLT)*, pages 355–366, 2008.

Pierre Gaillard, Gilles Stoltz, and Tim Van Erven. A second-order bound with excess losses. In *Proc Conference on Learning Theory (COLT)*, pages 176–196, 2014. URL https://arxiv.org/pdf/1402.2044.pdf.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

David P. Helmbold and Manfred K. Warmuth. Learning permutations with exponential weights. *Journal of Machine Learning Research*, 10(58):1705–1736, 2009. URL http://jmlr.org/papers/v10/helmbold09a.html.

Ruitong Huang, Tor Lattimore, András György, and Csaba Szepesvári. Following the leader and fast rates in linear prediction: curved constraint sets and other regularities. In *Advances in Neural Information Processing Systems*, pages 4970–4978, 2016.

Adam Tauman Kalai and Santosh Vempala. Efficient algorithms for on-line optimization. *Journal of Computer and System Sciences*, 71, 2016.

Wouter M Koolen and Tim Van Erven. Second-order quantile methods for experts and combinatorial games. In *Proc Conference on Learning Theory (COLT)*, pages 1155–1175, 2015.

Wouter M. Koolen, Peter Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4464–4472, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Cong Han Lim and Stephen J. Wright. Beyond the birkhoff polytope: Convex relaxations for vector permutation problems. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2168–2176, Cambridge, MA, USA, 2014. MIT Press.

Scott Linderman, Gonzalo Mena, Hal Cooper, Liam Paninski, and John Cunningham. Reparameterizing the birkhoff polytope for variational permutation inference. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1618–1627. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/linderman18a.html.

Mark Hickman, Pitu Mirchandani, Stefan Voß. *Computer-aided Systems in Public Transport*. Springer-Verlag, Berlin Heidelberg, 2008.

Martin Grötschel, László Lovász, Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*. Springer-Verlag, Berlin Heidelberg, 1993.

Jaouad Mourtada and Stéphane Gaïffas. On the optimality of the Hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20:1–28, 2019.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, 22(4):1679–1706, 10 1994. doi: 10.1214/aop/1176988477. URL https://doi.org/10.1214/aop/1176988477.

Daiki Suehiro, Kohei Hatano, Shuji Kijima, Eiji Takimoto, and Kiyohito Nagano. Online prediction under submodular constraints. In Nader H. Bshouty, Gilles Stoltz, Nicolas Vayatis, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 260–274, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-34106-9.

Tim van Erven and Wouter M Koolen. Metagrad: Multiple learning rates in online learning. In *Advances in Neural Information Processing Systems*, pages 3666–3674, 2016.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54): 1793–1861, 2015. URL http://jmlr.org/papers/v16/vanerven15a.html.

Manfred K Warmuth and Dima Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9:2287–2320, 2008.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Proc Conference on Learning Theory (COLT)*, pages 1–29, 2018.

Jun Zhang. Binary choice, subset choice, random utility, and ranking: A unified perspective using the permutahedron. *Journal of Mathematical Psychology*, 48(2):107–134, 2004. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2003.12.002. URL https://www.sciencedirect.com/science/article/pii/S0022249603001020.