# Minimizing Regret on Reflexive Banach Spaces and Nash Equilibria in Continuous Zero-Sum Games

**Authored by:**

Walid Krichene
Alexandre Bayen
Maximilian Balandat
Claire Tomlin

**Abstract**

We study a general adversarial online learning problem, in which we are given a decision set X' in a reflexive Banach space X and a sequence of reward vectors in the dual space of X. At each iteration, we choose an action from X', based on the observed sequence of previous rewards. Our goal is to minimize regret, defined as the gap between the realized reward and the reward of the best fixed action in hindsight. Using results from infinite dimensional convex analysis, we generalize the method of Dual Averaging (or Follow the Regularized Leader) to our setting and obtain upper bounds on the worst-case regret that generalize many previous results. Under the assumption of uniformly continuous rewards, we obtain explicit regret bounds in a setting where the decision set is the set of probability distributions on a compact metric space S. Importantly, we make no convexity assumptions on either the set S or the reward functions. We also prove a general lower bound on the worst-case regret for any online algorithm. We then apply these results to the problem of learning in repeated two-player zero-sum games on compact metric spaces. In doing so, we first prove that if both players play a Hannan-consistent strategy, then with probability 1 the empirical distributions of play weakly converge to the set of Nash equilibria of the game. We then show that, under mild assumptions, Dual Averaging on the (infinite-dimensional) space of probability distributions indeed achieves Hannan-consistency.

## 1 Paper Body

Regret analysis is a general technique for designing and analyzing algorithms for sequential decision problems in adversarial or stochastic settings (Shalev-Shwartz, 2012; Bubeck and Cesa-Bianchi, 2012). Online learning algorithms have applications in machine learning (Xiao, 2010), portfolio optimization (Cover, 1991), online convex optimization (Hazan et al., 2007) and other areas. Regret analysis also plays an important role in the study of repeated play of finite games (Hart and MasColell, 2001). It is well known, for example, that in a two-player zero-sum finite game, if both players play according to a Hannan-consistent strategy (Hannan, 1957), their (marginal) empirical distributions of play almost surely converge to the set of Nash equilibria of the game (Cesa-Bianchi and Lugosi, 2006). Moreover, it can be shown that playing a strategy that achieves sublinear regret almost surely guarantees Hannan-consistency. A natural question then is whether a similar result holds for games with infinite action sets. In this article we provide a positive answer. In particular, we prove that in a continuous two-player zero sum game over compact (not necessarily convex) metric spaces, if both players follow a Hannan-consistent strategy, then with probability 1 their empirical distributions of play weakly converge to the set of Nash equilibria of the game. This in turn raises another important question: Do algorithms that ensure Hannan-consistency exist in such a setting? More generally, can one develop algorithms that guarantee sub-linear growth of the worst-case regret? We answer these questions affirmatively as well. To this end, we develop a general framework to study the Dual Averaging (or Follow the Regularized Leader) method on reflexive Banach spaces. This framework generalizes a wide range of existing 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

results in the literature, including algorithms for online learning on finite sets (Arora et al., 2012) and finite-dimensional online convex optimization (Hazan et al., 2007). Given a convex subset X of a reflexive Banach space X, the generalized Dual Averaging (DA) method maximizes, at each iteration, the cumulative past rewards (which are elements of X ? , the dual space of X) minus a regularization term h. We show that under certain conditions, the maximizer in the DA update is the Fr?chet gradient Dh? of the regularizer?s conjugate function. In doing so, we develop a novel characterization of the duality between essential strong convexity of h and essential Fr?chet differentiability of h? in reflexive Banach spaces, which is of independent interest. We apply these general results to the problem of minimizing regret when the rewards are uniformly continuous functions over a compact metric space S. Importantly, we do not assume convexity of either S or the rewards, and show that it is possible to achieve sublinear regret under a mild geometric condition on S (namely, the existence of a locally Q-regular Borel measure). We provide explicit bounds for a class of regularizers, which guarantee sublinear worst-case regret. We also prove a general lower bound ?on the regret for any online algorithm and show that DA asymptotically achieves this bound up to a log t factor. Our results are related to work by Lehrer (2003) and Sridharan and Tewari (2010); Srebro et al. (2011). Lehrer (2003) gives necessary geometric conditions for Blackwell approachability in infinitedimensional spaces, but no implementable algorithm guaranteeing Hannan-consistency. Sridharan and Tewari (2010) derive general regret bounds for Mirror Descent (MD) under the assumption that the strategy set is uniformly bounded in the norm of the Banach space. We do not make such an assumption here. In fact, this assumption does not hold in general for our applications in Section 3. The paper is organized as follows: In Section 2 we introduce and provide a general analysis of Dual Averaging in reflexive Banach spaces. In Section 3 we apply these results to obtain explicit regret bounds on compact metric spaces with uniformly continuous reward functions. We use these results in Section 4 in the context of learning Nash equilibria in continuous two-player zero sum games, and provide a numerical example in Section 4. All proofs are given in the supplementary material.

2

Regret Minimization on Reflexive Banach Spaces

Consider a sequential decision problem in which we are to choose a sequence (x1 , x2 , . . . ) of actions from some feasible subset X of a reflexive Banach space X, and seek to maximize a sequence (u1 (x1 ), u2 (x2 ), . . . ) of rewards, where the u? : X ? R are elements of a given subset U ? X ? , with X ? the dual space of X. We assume that xt , the action chosen at time t, may only depend on the sequence of previously observed reward vectors (u1 , . . . , ut?1 ). We call any such algorithm an online algorithm. We consider the adversarial setting, i.e., we do not make any distributional assumptions on the rewards. In particular, they could be picked maliciously by some adversary. The notion of regret is a standard measure of performance for such a sequential decision problem. For a sequence (u1 , . . . , ut ) of reward vectors, and a sequence of decisions (x1 , . . . , xt ) produced by an algorithm, the regret of the algorithm w.r.t. a (fixed) ? X is the gap between the realized Pt decision xP t reward and the reward under x, i.e., Rt (x) := ? =1 u? (x) ? ? =1 u? (x? ). The regret is defined as Rt := supx?X Rt (x). An algorithm is said to have sublinear regret if for any sequence (ut )t?1 in the set of admissible reward functions U, the regret grows sublinearly, i.e. lim supt Rt /t ? 0.

Example 1. Consider a finite action set S = {1, . . . , n}, let X = X ? = Rn , and let X = ?n?1 , the probability simplex in Rn . A reward function can be identified with a vector u ? Rn , such that the i-th element ui is the reward of action i. A choice x ? X corresponds to a randomization over the n actions in S. This is the classic setting of many regret-minimizing algorithms in the literature. Example 2. Suppose S is a compact metric space with ? a finite measure on S. Consider X = X ? = L2 (S, ?) and let X = {x ? X : x ? 0 a.e., kxk1 = 1}. A reward function is an L2 integrable function on S, and each choice x ? X corresponds to a probability distribution (absolutely continuous w.r.t. ?) over S. We will explore a more general variant of this problem in Section 3. In this Section, we prove a general bound on the worst-case regret for DA. DA was introduced by Nesterov (2009) for (finite dimensional) convex optimization, and has also been applied to online learning, e.g. by Xiao (2010). In the finite dimensional case, the

Pt method solves, at each iteration, the optimization problem xt+1 = arg maxx?X ?t ? =1 u? , x ? h(x), where h is a strongly convex 2

regularizer defined on X ? Rn and (?t )t?0 is a sequence of learning rates. The regret analysis of the method relies on the duality between strong convexity and smoothness (Nesterov, 2009, Lemma 1). In order to generalize DA to our

Banach space setting, we develop an analogous duality result in Theorem 1. In particular, we show that the correct notion of strong convexity is (uniform) essential strong convexity. Equipped with this duality result, we analyze the regret of the Dual Averaging method and derive a general bound in Theorem 2. 2.1

Preliminaries

Let (X, k ? k) be a reflexive Banach space, and denote by h ? , ? i : X ? X ? ? R the canonical pairing between X and its dual space X ? , so that hx, ?i := ?(x) for all x ? X, ? ? X ? . By the effective domain of an extended real-valued function f : X ? [??, +?] we mean the set dom f = {x ? X : f (x) ¡ +?}. A function f is proper if f ¿ ?? and dom f is non-empty. The conjugate or Legendre-Fenchel transform of f is the function f ? : X ? ? [??, +?] given by f ? (?) = sup hx, ?i ? f (x) (1) x?X

for all ? ? X ? . If f is proper, lower semicontinuous and convex, its subdifferential ?f is the set-valued mapping ?f (x) = ? ? X ? : f (y) ? f (x) + hy ? x, ?i for all y ? X . We define dom ?f := {x ? X : ?f (x) 6= ?}. Let ? denote the set of all convex, lower semicontinuous functions ? : [0, ?) ? [0, ?] such that ?(0) = 0, and let

?U := ? ? ? : ?r ¿ 0, ?(r) ¿ 0 ?L := ? ? ? : ?(r)/r ? 0, as r ? 0 (2) We now introduce some definitions. Additional results are reviewed in the supplementary material. Definition 1 (Str?mberg, 2011). A proper convex lower semicontinuous function f : X ? (??, ?] is essentially strongly convex if (i) f is strictly convex on every convex subset of dom ?f (ii) (?f )?1 is locally bounded on its domain (iii) for every x0 ? dom ?f there exists ?0 ? X ? and ? ? ?U such that f (x) ? f (x0 ) + hx ? x0 , ?0 i + ?(kx ? x0 k), ?x ? X.

(3)

If (3) holds with ? independent of x0 , f is uniformly essentially strongly convex with modulus ?. Definition 2 (Str?mberg, 2011). A proper convex lower semicontinuous function f : X ? (??, ?] is essentially Fr?chet differentiable if int dom f 6= ?, f is Fr?chet differentiable on int dom f with Fr?chet derivative Df , and kDf (xj )k? ? ? for any sequence (xj )j in int dom f converging to some boundary point of dom f . Definition 3. A proper Fr?chet differentiable function f : X ? (??, ?] is essentially strongly smooth if ? x0 ? dom ?f, ? ?0 ? X ? , ? ? ?L such that f (x) ? f (x0 ) + h?0 , x ? x0 i + ?(kx ? x0 k), ? x ? X. (4) If (4) holds with ? independent of x0 , f is uniformly essentially strongly smooth with modulus ?. With this we are now ready to give our main duality result: Theorem 1. Let f : X ? (??, +?] be proper, lower semicontinuous and uniformly essentially strongly convex with modulus ? ? ?U . Then (i) f ? is proper and essentially Fr?chet differentiable with Fr?chet derivative Df ? (?) = arg max hx, ?i ? f (x).

(5)

x?X

If, in addition, ?? (r) := ?(r)/r is strictly increasing, then ?

kDf ? (?1 ) ? Df ? (?2 )k ? ?? ?1 k?1 ? ?2 k? /2 .

In other words, Df is uniformly continuous with modulus of continuity ?(r) = ?? ?

(6) ?1
(r/2).
?
(ii) f is uniformly essentially smooth with modulus ? . 1/?
Corollary 1. If ?(r) ? C r1+? , ? r ? 0 then kDf ? (?1 ) ? Df ? (?2 )k ? (2C)?1/? k?1 ? ?2 k? . 2 In particular, with ?(r) = K 2 r , Definition 1 becomes the classic definition of K-strong convexity, and (6) yields the result familiar from the finite-dimensional case that the gradient Df ? is 1/K Lipschitz with respect to the dual norm (Nesterov, 2009, Lemma 1). 3
2.2
Dual Averaging in Reflexive Banach Spaces
We call a proper convex function h : X ? (??, +?] a regularizer function on a set X ? X if h is essentially strongly convex and dom h = X . We emphasize that we do not assume h to be Fr?chet-differentiable. Definition 1 in conjunction with Lemma S.1 (supplemental material) implies that for any regularizer h, the supremum of any function of the form h ? , ?i ? h( ? ) over X, where ? ? X ? , will be attained at a unique element of X , namely Dh? (?), the Fr?chet gradient of h? at ?.

DA with regularizer h and a sequence of learning rates (?t )t?1 generates a sequence of decisions Pt using the simple update rule xt+1 = Dh? (?t Ut ), where Ut = ? =1 u? and U0 := 0. Theorem 2. Let h be a uniformly essentially strongly convex regularizer on X with modulus ? and let (?t )t?1 be a positive non-increasing sequence of learning rates. Then, for any sequence of payoff functions (ut )t?1 in X ? for which there exists M ¡ ? such that supx?X —hut , xi— ? M for all t, the sequence of plays (xt )t?0 given by
Pt xt+1 = Dh? ?t ? =1 u? (7) ensures that t t t ?
X X h(x) ? h X ? ?1 Rt (x) := hu? , xi ? hu? , x? i ? + ku? k? ?? ?1 ku? k? (8) ?t 2 ? =1 ? =1 ? =1 where h = inf x?X h(x), ?? (r) := ?(r)/r and ?0 := ?1 . It is possible to obtain a regret bound similar to (8) also in a continuous-time setting. In fact, following Kwon and Mertikopoulos (2014), we derive the bound (8) by first proving a bound on a suitably defined notion of continuous-time regret, and then bounding the difference between the continuous-time and discrete-time regrets. This analysis is detailed in the supplementary material. Note that the condition that supx?X —hut , xi— ? M in Theorem 2 is weaker than the one in Sridharan and Tewari (2010), as it does not imply a uniformly bounded strategy set (e.g., if X = L2 (R) and X is the set of distributions on X, then X is unbounded in L2 , but the condition may still hold).

Theorem 2 provides a regret bound for a particular choice x ? X . Recall that Rt := supx?X Rt (x). In Example 1 the set X is compact, so any continuous regularizer h will be bounded, and hence taking the supremum over x in (8) poses no issue. However, this is not the case in our general setting, as the regularizer may R be unbounded on X . For instance, consider Example 2 with the entropy regularizer h(x) = S x(s) log(x(s))ds, which is easily seen to be unbounded on X . As a consequence, obtaining a worst-case bound will in general require additional assumptions on the reward functions and the decision set X . This will be investigated in detail in Section 3. Corollary 2. Suppose that ?(r) ? C r1+? , ? r ? 0 for some C ¿ 0 and ? ¿ 0. Then t X h(x) ? h 1/? 1+1/? Rt (x) ? + (2C)?1/? ?? ?1 ku? k? . (9) ?t ? =1 In particular, if kut k? ? M for all t and ?t = ? t?? , then ? ? 1/? 1+1/? 1??/? h(x) ? h ? t + M t . Rt (x) ? ? ? ? ? 2C

(10)

?

Assuming h is bounded, optimizing over ? yields a rate of Rt (x) = O(t 1+? ). In particular, ? if K 2 ?(r) = 2 r , which corresponds to the classic definition of strong convexity, then Rt (x) = O( t). For non-vanishing u? we will need that ?t & 0 for the sum in (9) to converge. Thus we could get potentially tighter control over the rate of this term for ? ¡ 1, at the expense of larger constants.

3

Online Optimization on Compact Metric Spaces

We now apply the above results to the problem minimizing regret on compact metric spaces under the additional assumption of uniformly continuous reward functions. We make no assumptions on convexity of either the feasible set or the rewards. Essentially, we lift the non-convex problem of minimizing a sequence of functions over the (possibly non-convex) set S to the convex (albeit infinitedimensional) problem of minimizing a sequence of linear functionals over a set X of probability measures (a convex subset of the vector space of measures on S). 4

3.1

An Upper Bound on the Worst-Case Regret

Let (S, d) be a compact metric space, and let ? be a Borel measure on S. Suppose that the reward vectors u? are given by elements in Lq (S, ?), where q ¿ 1. Let X = Lp (S, ?), where p and q are H?lder conjugates, i.e., p1 + 1q = 1. Consider X = {x ? X : x ? 0 a.e., kxk1 = 1}, the set of probability measures on S that are absolutely continuous w.r.t. ? with p-integrable Radon-Nikodym derivatives. Moreover, denote by Z the class of non-decreasing ? : [0, ?) ? [0, ?] such that limr?0 ?(r) = ?(0) = 0. The following assumption will be made throughout this section: Assumption 1. The reward vectors ut have modulus of continuity ? on S, uniformly in t. That is, there exists ? ? Z such that —ut (s) ? ut (s0 )— ? ?(d(s, s0 )) for all t and for all s, s0 ? S. Let B(s, r) = {s0 ? S : d(s, s0 ) ¡ r} and denote by B(s, ?) ? X the elements of X with support contained in B(s, ?). Furthermore, let DS := sups,s0 ?S d(s, s0 ). Then we have the following: Theorem 3. Let (S, d) be compact, and suppose that Assumption 1 holds. Let h be a uniformly essentially strongly convex regularizer on X with modulus ?, and let (?t )t?1 be a positive nonincreasing sequence of learning rates. Then, under (7), for any positive sequence (?t )t?1 , t ?

X sups?S inf x?B(s,?t ) h(x) ? h ? ?1 Rt ? + t ?(?t ) + ku? k? ?? ?1 ku? k? . (11) ?t 2 ? =1 Remark 1. The sequence (?t )t?1 in Theorem 3 is not a parameter of the algorithm, but rather a parameter in the regret bound. In particular, (11) holds true for any such sequence, and we will use this fact later on to obtain explicit bounds by instantiating (11) with a particular choice of (?t )t?1 . It is important to realize that the infimum over B(s, ?t ) in (11) may be infinite, in which case the bound is meaningless. This happens for example if s is an isolated point of some S ? Rn and ? is the Lebesgue measure, in which case B(s, ?t ) = ?. However, under an additional regularity assumption on the measure ? we can avoid such degenerate situations. Definition 4 (Heinonen. et al., 2015). A Borel measure ? on a metric space (S, d) is (Ahlfors) Q-regular if there exist 0 ¡ c0 ? C0 ¡ ? such that for any open ball B(s, r) c0 rQ ? ?(B(s, r)) ? C0 rQ . (12) We say that ? is r0 -locally Q-regular if (12) holds for all 0 ¡ r ? r0 . Intuitively, under an r0 -locally Q-regular measure, the mass in the neighborhood of any point of S is uniformly bounded from above and below. This will allow, at each iteration t, to assign sufficient probability mass around the maximizer(s) of the cumulative reward function. Example 3. The canonical example for a Q-regular measure is the Lebesgue measure ? on Rn . If d is the metric induced by the Euclidean norm, then Q = n and the bound (12) is tight with c0 = C0 , a dimensional constant. However, for general sets S ? Rn , ? need not be locally Q-regular. A sufficient condition for local regularity of ? is that S is v-uniformly fat (Krichene et al., 2015). Assumption 2. The measure ? is r0 -locally Q-regular on (S, d). Under Assumption 2, B(s, ?t ) 6= ? for all s ? S and ?t ¿ 0, hence we may hope for a bound on inf x?B(s,?t ) h(x) uniform in s. To obtain explicit convergence rates, we have to consider a more specific class of regularizers. 3.2

Explicit Rates for f -Divergences on Lp (S)

We consider a particular class of regularizers called f -divergences or Csisz?r divergences (Csisz?r, 1967). Following Audibert et al. (2014), we define ?-potentials and the associated f -divergence. Definition 5. Let ? ? 0 and a ? (??, +?]. A continuous increasing diffeomorphism ? : (??, a) ? (?, ?), is an ?-potential if limz??? ?(z) = ?, limz?a ?(z) R= +? and x ?1 ?(0) ? 1. Associated to ? is the convex function R f? : [0, ?) ? R defined by f? (x) = 1 ? (z) dz and the f? -divergence, defined by h? (x) = S f? x(s) d?(s) + ?X (x), where ?X is the indicator function of X (i.e. ?X (x) = 0 if x ? X and ?X (x) = +? if x ? / X ). A remarkable fact is that for regularizers based on ? potentials, the DA update (7) can be computed efficiently. More precisely, it can be shown (see Proposition 3 in Krichene (2015)) that the maximizer in this case has aPsimple expression in terms of the dual problem, and the problem of computing t xt+1 = Dh? (?t ? =1 u? ) reduces to computing a scalar dual variable ?t? . 5

Proposition 1. Suppose that ?(S) = 1, and that Assumption 2 holds with constants r0 ¿ 0 and 0 ¡ c0 ? C0 ¡ ?. Under the Assumptions of Theorem 3, with h = h? the regularizer associated to an ?-potential ?, we have that, for any positive sequence (?t )t?1 with ?t ? r0 , t

Rt min(C0 ?Q 1X ?Q t , ?(S)) ?1 ?? ?1 ? f? c?1 ? + ?(? ) + ku k ? ? ku k (13) t ? ? ? ? . t 0 t t ?t t ? =1 2 For particular choices of the sequences (?t )t?1 and (?t )t?1 , we can derive explicit regret rates. 3.3

Analysis for Entropy Dual Averaging (The Generalized Hedge Algorithm) Rx Taking ?(z) = ez?1 , we have that f? (x) = 1 ??1 (z)dz = x log x, and hence the regularizer is R exp ?(s) h? (x) = S x(s) log x(s)d?(s). Then Dh? (?)(s) = k exp ?(s)k1 . This corresponds to a generalized Hedge algorithm (Arora et al., 2012; Krichene et al., 2015) or the entropic barrier of Bubeck and Eldan (2014) for Euclidean spaces. The regularizer h? can be shown to be essentially strongly convex with modulus ?(r) = 12 r2 . Corollary 3. Suppose that ?(S) = 1, that ? is r0 -locally Q-regular with constants c0 , C0 , that kut k? ? M for all t, and that ?(r) = C? r? for 0 ¡ ? ? 1 (that is, theprewards are ?-H?lder continuous). Then, under Entropy Dual Averaging, choosing ?t = ? log t/t with

Q 1/2 ?1 ?Q/? 1 C0 Q ?=M ) + 2? and ? ¿ 0, we have that 2c0 log(c0 ? r

r Rt 2C0 log t Q ?Q/? ) + ? 2M log(c?1 ? + C ? (14) ? 0 t c0 2? t p whenever log t/t ¡ r0? ??1 . One can now further optimize over the choice of ? to obtain the best constant in the bound. Note also that the case ? = 1 corresponds to Lipschitz continuity. 3.4

A General Lower Bound

Theorem 4. Let (S, d) be compact, suppose that Assumption 2 holds, and let w : R ? R be any function with modulus of continuity ? ? Z such that kw(d( ? , s0 ))kq ? M for some s0 ? S for which there exists s ? S with d(s, s0 ) = DS . Then for any online algorithm, there exist a sequence (u? )t? =1 of reward vectors u? ? X ? with ku? k? ? M and modulus of continuity ?? ¡ ? such that w(DS ) ? Rt ? ? t, (15) 2 2 Maximizing the constant in (15) is of interest in order to benchmark the bound against the upper bounds obtained in the previous sections. This problem is however quite challenging, and we will defer this analysis to future work. For H?lder-continuous functions, we have the following result: Proposition 2. In the setting of Theorem 4, suppose that ?(S) = 1 and that ?(r) = C? r? for some 0 ¡ ? ? 1. Then

1/? min C? DS? , M ? ? Rt ? t. (16) 2 2 ? Observe that, up to a log t factor, the asymptotic rate of this general lower bound for any online algorithm matches that of the upper bound (14) of Entropy Dual Averaging.

4

Learning in Continuous Two-Player Zero-Sum Games

Consider a two-player zero sum game G = (S1 , S2 , u), in which the strategy spaces S1 and S2 of player 1 and 2, respectively, are Hausdorff spaces, and u : S1 ? S2 ? R is the payoff function of player 1 (as G is zero-sum, the payoff function of player 2 is ?u). For each i, denote by Pi := P(Si ) the set of Borel probability measures on Si . Denote S := S1 ? S2 and P := P1 ? P2 . For a (joint) mixed strategy x ? P, we define the natural extension u ? : P ? R by u ?(x) := Ex [u] = R u(s1 , s2 ) dx(s1 , s2 ), which is the expected payoff of player 1 under x. S 6

A continuous zero-sum game G is said to have value V if ?(x1 , x2 ) = V. ?(x1 , x2 ) = inf sup u sup inf u x1 ?P1 x2 ?P2

(17)

x2 ?P2 x1 ?P1

The elements x1 ? x2 ? P at which (17) holds are the (mixed) Nash Equilibria of G. We denote the set of Nash equilibria of G by N (G). In the case of finite games, it is well known that every two-player zero-sum game has a value. This is not true in general for continuous games, and additional conditions on strategy sets and payoffs are required, see e.g. (Glicksberg, 1950). 4.1

Repeated Play

We consider repeated play of the continuous two-player zero-sum game. Given a game G and a sequence of plays (s1t )t?1 and (s2t )t?1 , we say that player i has sublinear (realized) regret if

t t X X 1 i ?i i ?i lim sup sup ui (s , s? ) ? ui (s? , s? ) ? 0 (18) t?? t si ?Si ? =1 ? =1 where we use ?i to denote the other player.

A strategy ? i for player i is, loosely speaking, a (possibly random) mapping from past observations to its actions. Of primary interest to us are Hannan-consistent strategies: Definition 6 (Hannan, 1957). A strategy ? i of player i is Hannan consistent if, for any sequence (st?i )t?1 , the sequence of plays (sti )t?1 generated by ? i has sublinear regret almost surely. Note that the almost sure statement in Definition 6 is with respect to the randomness in the strategy ? i . The following result is a generalization of its counterpart for discrete games (e.g. Corollary 7.1 in (Cesa-Bianchi and Lugosi, 2006)): Proposition 3. Suppose G has value V and consider a sequence of plays (s1t )t?1 , (s2t )t?1 and Pt assume that both players have sublinear realized regret. Then limt?? 1t ? =1 u(s1? , s2? ) = V . As in the discrete case (Cesa-Bianchi and Lugosi, 2006), we can also say something about convergence of the empirical distributions of play to the set of Nash Equilibria. Since these distributions have finite support for every t, we can at best hope for convergence in the weak sense as follows: Theorem 5. Suppose that in a repeated two-player zero sum game Pt G that has a value both players follow a Hannan-consistent strategy, and denote by x ?it = 1t ? =1 ?si? the marginal empirical distribution of play of player i at iteration t. Let x ?t := (? x1t , x ?2t ). Then x ?t * N (G) almost surely, that is, with probability 1 the sequence (? xt )t?1 weakly converges to the set of Nash equilibria of G. Corollary 4. If G has a unique Nash equilibrium x? , then with probability 1, x ?t * x? . 4.2

Hannan-Consistent Strategies

By Theorem 5, if each player follows a Hannan-consistent strategy, then the empirical distributions of play weakly converge to the set of Nash equilibria of the game. But do such strategies exist? Regret minimizing strategies are intuitive candidates, and the intimate connection between regret minimization and learning in games is well studied in many cases, e.g. for finite games (CesaBianchi and Lugosi, 2006) or potential games (Monderer and Shapley, 1996). Using our results from Section 3, we will show that, under the appropriate assumption on the information revealed to the player, no-regret learning based on Dual Averaging leads to Hannan consistency in our setting. Specifically, suppose that after each iteration t, each player i observes a partial payoff function u ?it : Si ? R describing their payoff as a function of only their own action, si , holding the action played by the other player fixed. That is, u ?1t (s1 ) := u(s1 , s2t ) and u ?2t (s2 ) := ?u(s1t , s2 ). Remark 2. Note that we do not assume that the players have knowledge of the joint utility function u. However, we do assume that the player has full information feedback, in the sense that they observe partial reward functions u( ? , s?i ? ) on their entire action set, as opposed to only observing the reward u(s1? , s2? ) of the action played (the latter corresponds to the bandit setting). ?ti = (? ui? )t? =1 the sequence of partial payoff functions observed by player i. We use We denote by U i Ut to denote the set of all possible such histories, and define U0i := ?. A strategy ? i of player i is a i i collection (?ti )? t=1 of (possibly random) mappings ?t : Ut?1 ? Si , such that at iteration t, player i i plays sit = ?ti (Ut?1 ). We make the following assumption on the payoff function: 7

Assumption 3. The payoff function u is uniformly continuous in si with modulus of continuity independent of s?i for i = 1, 2. That is, for each i there exists ?i ? Z such that —u(s, s?i ) ? u(s0 , s?i )— ? ?i (di (s, s0 )) for all s?i ? S?i . It is easy to see that Assumption 3 implies that the game has a value (see supplementary material). It also makes our setting compatible with that of Section 3. Suppose now that each player randomizes their play according to the sequence of probability distributions on Si generated by DA with regularizer hi . That is, suppose that each ?ti is a random variable with the following distribution: Pt?1 i ?ti ? Dh?i ?t?1 ? =1 u ?? . (19) Theorem 6. Suppose that player i uses strategy ? i according to (19), and that the DA algorithm ensures sublinear regret (i.e. lim supt Rt /t ? 0). Then ? i is Hannan-consistent. Corollary 5. If both players use strategies according to (19) with the respective Dual Averaging ensuring that lim supt Rt /t ? 0, then with probability 1 the sequence (? xt )t?1 of empirical distributions of play weakly converges to the set of Nash equilibria of G. Example Consider a zero-sum game G1 between two players on the unit interval with payoff func1 2 tion u(s1 , s2 ) = s1 s2 ? a1 s1 ? a2 s2 , where a1 = e?2 e?1 and a = e?1 . It is easy to verify that the pair

exp(1?s) x1 , x2 = exp(s) is a mixed-strategy Nash equilibrium of G1 . For sequences (s1? )t? =1 e?1 , e?1 2 t and (s? )? =1 , the cumulative payoff functions for fixed action s ? [0, 1] are given, respectively, by

Ut1 (s1 ) = ?t? =1 s2? ? a1 t s1 ? a2 ?t? =1 s2? Ut2 (s2 ) = a2 t ? ?t? =1 s1? s2 ? a1 ?t? =1 s1? If each player i uses the Generalized Hedge Algorithm with learning rates (?? )t? =1 , their strategy in period t is to sample from the distribution xit (s) ? exp(?ti s), where ?t1 = ?t (?t? =1 s2? ? a1 t) and ?t2 = ?t (a2 t ? ?t? =1 s1? ). Interestingly, in this case the sum of the opponent?s past plays is a sufficient statistic, in the sense that it completely determines the mixed strategy at time t. 2.5 2.0 1.5 1.0 0.5 0.0 2.5 2.0 1.5 1.0 0.5 0.0 0.0

player 1, t=5000

player 1, t=50000

x1 (s)

player 1, t=500000

x1 (s)

player 2, t=5000

x1 (s)

player 2, t=50000

player 2, t=500000

x2 (s)

0.2

0.4

0.6

0.8

x2 (s)

1.0

0.0

0.2
0.4
0.6
0.8
x2 (s)
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Figure 1: Normalized histograms of the empirical distributions of play in G (100 bins) Figure 1 shows normalized histograms of the empirical distributions of play at different iterations t. As t grows the histograms approach the equilibrium densities x1 and x2 , respectively. However, this does not mean that the individual strategies xit converge. Indeed, Figure 2 shows the ?ti oscillating around the equilibrium parameters 1 and ?1, respectively, even for very large t. We do, however, observe that the time-averaged parameters ? ? ti converge to the equilibrium values 1 and ?1. 2

?t1 ? ? t1
?t2 ? ? t2
1 0 ?1 100
101
102
103
104

Figure 2: Evolution of parameters ?ti and ? ? ti :=
105
Pt 1 t
? =1
106
??i in G1

In the supplementary material we provide additional numerical examples, including one that illustrates how our algorithms can be utilized as a tool to compute approximate Nash equilibria in continuous zero-sum games on nonconvex domains. 8

## 2 References

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a metaalgorithm and applications. Theory of Computing, 8(1):121?164, 2012. Jean-Yves Audibert, S?bastien Bubeck, and G?bor Lugosi. Regret in online combinatorial optimization. Mathematics of Operations Research, 39(1):31?45, 2014. S. Bubeck and R. Eldan. The entropic barrier: a simple and optimal universal self-concordant barrier. ArXiv e-prints, December 2014. S?bastien Bubeck and Nicol? Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multiarmed bandit problems. Foundations and Trends in Machine Learning, 5(1):1?122, 2012. Nicolo Cesa-Bianchi and Gabor Lugosi. Prediction, Learning, and Games. Cambridge UP, 2006. Thomas M. Cover. Universal portfolios. Mathematical Finance, 1(1):1?29, 1991. Imre Csisz?r. Information-type measures of difference of probability distributions and indirect observations. Studia Scientiarum Mathematicarum Hungarica, 2:299?318, 1967. Irving L. Glicksberg. Minimax theorem for upper and lower semicontinuous payoffs. Research Memorandum RM-478, The RAND Corporation, Oct 1950. James Hannan. Approximation to Bayes risk in repeated play. In Contributions to the Theory of Games, vol III of Annals of Mathematics Studies 39. Princeton University Press, 1957. Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. Journal of Economic Theory, 98(1):26 ? 54, 2001. Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. Machine Learning, 69(2-3):169?192, 2007. Juha Heinonen., Pekka Koskela, Nageswari Shanmugalingam, and Jeremy T. Tyson. Sobolev Spaces on Metric Measure Spaces: An Approach Based on Upper Gradients. New Mathematical Monographs. Cambridge University Press, 2015. Walid Krichene. Dual averaging on compactly-supported distributions and application to no-regret learning on a continuum. CoRR, abs/1504.07720, 2015. Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. The Hedge Algorithm on a Continuum. In 32nd International Conference on Machine Learning, pages 824?832, 2015. Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. ArXiv e-prints, January 2014. Ehud Lehrer. Approachability in infinite dimensional spaces. International Journal of Game Theory, 31(2):253?268, 2003. Dov Monderer and Lloyd S. Shapley. Potential games. Games and Economic Behavior, 14(1):124 ? 143, 1996. Yurii Nesterov. Primal-dual subgradient methods for convex problems. Mathematical Programming, 120(1):221?259, 2009. Shai Shalev-Shwartz. Online learning and online convex optimization. Foundations and Trends in Machine Learning, 4(2):107?194, 2012. Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In Advances in Neural Information Processing Systems 24 (NIPS), pages 2645?2653. 2011. Karthik Sridharan and Ambuj Tewari. Convex games in banach spaces. In COLT 2010 - The 23rd Conference on Learning Theory,, pages 1?13, Haifa, Israel, June 2010. Thomas Str?mberg. Duality between Fr?chet differentiability and strong convexity. Positivity, 15(3): 527?536, 2011. Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res., 11:2543?2596, December 2010. 9