Strategies for convex potential games and an application to decision-theoretic online learning

Yoav Freund

June 21, 2021

Abstract

The backwards induction method due to Bellman [1] is a popular approach in optimiztion, optimal control, and many other areas of applied math. In this paper we analyze the backwords induction approach, under min/max conditions. We show that if the value function is has strictly positive derivatives of order 1-4 then the optimal strategy for the adversary is brownian motion. Using that fact we analyze different potential functions and show that the Normal-Hedge potential is optimal.

1 Introduction

Our analysis is an application of Bellman's equation [1] and the more recent work on drifting games by Schapire[5]. This setup has been studied in the past (BW algorithm, drifting games).

We describe a set of strategies for the decision-theoretic online learning problem (DTOL) []. To develop those strategies we replace discrete time: $t = 0, 1, 2, \ldots, T$ with continuous time: $t \in [0, T]$. Suppose there are N actions whose per iteration loss is in the range $[0, 1]$, and suppose that the game is repeated T times. There are several algorithm whose regret is bounded by $C\sqrt{T \ln N}$ for some constant C. Next Suppose that the actual losses are in the range $[0, 1/2]$ which would imply that the regret is at most $\frac{C}{2}\sqrt{T \ln N}$ and that the bound is twice large than it should be. We will show that a natural way to remove this slack is to say that time increases in increments of $1/4$ so that the final time is $T/4$. In this paper we show that this idea is very general and leads to a a min/max optimal algorithm for DTOL.

2 Setup

The game takes place on the set $(t, R) \in [0, T] \times \mathbb{R}$, where t corresponds to time, T correspond to the end time of the game (i.e. this is a bounded horizon game where the horizon is known to both players. R curresponds to (total) regret.

We will first consider the standard setting where time corresponds to the natural and is equal to the iteration number $t_i = i$. Later we expand the game the t_i can take on any real value in $[0, T]$ under the ccondition that $t_{i-1} \leq t_i \leq t_{i-1} + 1$.

The *state* of the game at time t_i is a distribution over regrets $\mathbb{R}$ denoted by $R \sim \mathbf{\Psi}(t_i)$. The initial distribution $\mathbf{\Psi}(0)$ is delta-function at $R = 0$. States $\mathbf{\Psi}(t_i)$ for $t > 0$ also correspond to distributions over regret. The state $\mathbf{\Psi}(t)$ is defined by $\mathbf{\Psi}(t_{i-1})$ and the choices made by the two players as described in the next section.

There are two players: the *learner* and the *adversary*. There is a value function defined for the final state $\Phi_{\mathrm{T}} = \Phi(T) : \mathbb{R} \to \mathbb{R}$.

The final potential $\Phi_{\mathrm{T}}(R)$ is fixed a-priori and is known to both players. $\Phi_{\mathrm{T}}(R)$ is restricted to be continuous, monotone non-decreasing and strictly convex. We will refer to these conditions as **CIC**.

The outcome of the game is the expected final value,

$$\phi(T) = \mathbf{E}_{R \sim \mathbf{\Psi}(T)} \left[\Phi_{\mathrm{T}}(R) \right] \tag{1}$$

The goal of the learner is to minimize $\phi(T)$ and the goal of the adversary is to maximize it.

3 Integer time game

We start with the standard setup in which time corresponds to the natural numbers, $t_i = i$. to simplify notation we will use the iteration number i instead of time.

Connecting this back to decision theoretic online learning (DTOL []). The state $\mathbf{\Psi}(i)$ corresponds to the distribution over the regret values of the experts on iteration i. Note that here the set of experts is allowed to be be uncountably infinite. In particular the adversary can assign to the experts with regret x at iteration t an arbitrary distribution of losses in the range $[-1, +1]$.

The game is defined by three parameters:

- T : The number of iterations
- $\Phi_{\mathrm{T}}(R)$: The final value function that is CIC.
- $0 < c \leq 1$ - An upper bound on aggregate loss (loss of the master) in a single iteration. Note that the cumulative aggregate loss is at most cT. Note that $c = 1$ is always satisfied and nullifies the constraint.

The transition from $\mathbf{\Psi}(i)$ to $\mathbf{\Psi}(i+1)$ is defined by the choices made by the adversary and the learner.

1. The learner chooses weights. Formally, this is a distribution over $R \in \mathbb{R}$: $P(i)$.
2. The adversary chooses the losses of the actions. Formally this is a mapping from $\mathbb{R}$ to distributions over $[-1, +1]$: $Q(i) : \mathbb{R} \to \Delta^{[-1,+1]}$. We use $l \sim Q(i, R)$ to denote the distribution over the instantanous loss associated with iteration i and regret R.
3. The aggregate loss (also called "the loss of the master") is calculated:

$$\ell(i) = \mathbf{E}_{R \sim \mathbf{\Psi}(i)} \left[P(i, R) \mathbf{E}_{l \sim Q(i,R)} [l] \right] \tag{2}$$

   The adversary is constrained to Q such $|\ell(i)| \leq c$. We define the *bias* at (i, R) to be $B(i, R) \doteq \mathbf{E}_{l \sim Q(i,R)} [l]$ which allows us to rewrite Eqn (2) as

$$\ell(i) = \mathbf{E}_{R \sim \mathbf{\Psi}(i)} [P(i, R) B(i, R)] \tag{3}$$

   note that $B(i, R)$ is in $[-1, 1]$ and that $\ell(i)$ is the mean of $B(i, \cdot)$. Note also that $-1-c \leq y-\ell(i) \leq 1+c$ corresponds to the instantanous regret. In the integer game, setting $c \geq 1$ is equivalent to placing no restriction on $\ell(i)$.
4. The state is updated.

$$\mathbf{\Psi}(i+1) = \mathbf{E}_{R \sim \mathbf{\Psi}(i)} [R \oplus Q(i, R)] \ominus \ell(i) \tag{4}$$

   Where $Q(i, R)$ is the distribution of the losses of experts that are at location R after iteration $i-1$. $R \oplus Q(i, R)$ is the same distribution shiften right by the amount R. $\mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} [\cdot]$ indicates the expectation over distribution, yealding a new distribution. Finally $\ominus \ell(i)$ is a shift left of the resulting distribtion according to the aggregate loss.

The final outcome of the game is given in Equation (1).

3.1 Strategies for integer time game

Consider the states at two consecutive iterations $\mathbf{\Psi}(i-1), \mathbf{\Psi}(i)$. Suppose that $\Phi(i,R)$, the value function for iteration i is fixed. We say that $\Phi^{\uparrow}(i-1,R)$ is a lower bound on the value at iteration $i-1$ if there is exists an adversarial strategy Q^* such that for any strategy of the learner P,

$$\mathbf{E}_{R\sim\mathbf{\Psi}(i-1)}\left[\Phi^{\uparrow}(i-1)\right] \leq \mathbf{E}_{R\sim\mathbf{\Psi}(i)}\left[\Phi(i)\right]$$

Similarly, $\Phi^{\downarrow}(i-1,R)$ is an upper potential if there exists a learner strategy P^* such that for any adversarial strategy Q,

$$\mathbf{E}_{R\sim\mathbf{\Psi}(i-1)}\left[\Phi^{\downarrow}(i-1)\right] \geq \mathbf{E}_{R\sim\mathbf{\Psi}(i)}\left[\Phi(i)\right]$$

Lemma 1. *Suppose $\Phi^{\uparrow}(i,R)$ is strictly convex with respect to R. Let $\mathcal{Q}(i-1,R,B)$ be the set of adversarial strategies $Q(i-1,R)$ that have bias $B = B(i-1,R) = \mathbf{E}_{y\sim Q(i-1,R)}[y]$ then the strategy in $\mathcal{Q}(i-1,R,B)$ that is best for the adversary is*

$$Q^p(i-1,R) = \begin{cases} +1 & w.p.\ \frac{1+B}{2} \\ -1 & w.p.\ \frac{1-B}{2} \end{cases} \tag{5}$$

$$\Phi^{\uparrow}(i-1,R) = p\Phi^{\uparrow}(i,R+1) + (1-p)\Phi^{\uparrow}(i,R-1) \tag{6}$$

which is strictly higher than $\Phi^{\uparrow}(i-1,R)$ for any other distribution in $\mathcal{Q}(i-1,R,B)$. In addition, $\Phi^{\uparrow}(i-1,R)$ is strictly convex.

In the next lemma we describe strategies for the adversary and the learner and prove upper and lower bounds on the potential that they guarantee.

Lemma 2. *If $\Phi(i,R)$ is CIC then*

1. *The adversarial strategy*

$$Q^{1/2}(i-1,R) = \begin{cases} -1 & w.p.\ 1/2 \\ +1 & w.p.\ 1/2 \end{cases} \tag{7}$$

Guarantees the lower potential

$$\Phi^{\uparrow}(i-1,R) = \frac{\Phi(i,R+1)+\Phi(i,R-1)}{2} \tag{8}$$

2. *The learner strategy:*

$$P^1(i-1,R) = \frac{1}{Z}\frac{\Phi(i,R+1+c)-\Phi(i,R-1-c)}{2} \tag{9}$$

Where Z is a normalization factor

$$Z = \mathbf{E}_{R\sim\mathbf{\Psi}(i)}\left[\frac{\Phi(i,R+1+c)-\Phi(i,R-1-c)}{2}\right]$$

guarantees the upper potential

$$\Phi^{\downarrow}(i-1,R) = \frac{\Phi(i,R+1+c)+\Phi(i,R-1-c)}{2} \tag{10}$$

Proof.

1. By symmetry adversarial strategy (7) guarantees that the aggregate loss (3) is zero regardless of the choice of the learner: $\ell(t) = 0$. Therefor the state update (4) is equivalent to the symmetric random walk:

$$\mathbf{\Psi}(i) = \frac{1}{2}((\mathbf{\Psi}(i-1) \oplus 1) + (\mathbf{\Psi}(i-1) \ominus 1))$$

Which in turn implies that if the adversary plays Q^* and the learner plays an arbitrary strategy P

$$\Phi^{\uparrow}(i-1, R) = \frac{1}{2}(\Phi(i, R-1) + \Phi(i, R+1)) \tag{11}$$

As this adversaril strategy is oblivious to the strategy, it guarantees that the average value at iteration i is *equal* to the average of the lower value at iteration $i-1$.

2. Plugging learner's strategy (9) into equation (3) we find that

$$\ell(i-1) = \frac{1}{Z_{i-1}} \mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} \left[(\Phi(i, R+1+c) - \Phi(i, R-1-c)) B(i-1, R) \right] \tag{12}$$

Consider the average value at iteration $i-1$ when the learner's strategy is P^* and the adversarial strategy is arbitrary Q:

$$\phi_{P^*,Q}(i-1, R) = \mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} \left[\mathbf{E}_{y \sim Q(i-1)(R)} \left[\Phi(i, R + y - \ell(i-1)) \right] \right] \tag{13}$$

As $\Phi(i, \cdot)$ is convex and as $(y - \ell(i-1)) \in [-1-c, 1+c]$,

$$\Phi(i, R+y) \leq \frac{\Phi(i, R+1+c) + \Phi(i, R-1-c)}{2} + (y - \ell(i)) \frac{\Phi(i, R+1+c) - \Phi(i, R-1-c)}{2} \tag{14}$$

Combining the equations (12) and (13) we find that

$$\begin{aligned}
\phi_{P^*,Q}(i-1, R) &= \mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} \left[\mathbf{E}_{y \sim Q(i-1)(R)} \left[\Phi(i, R + y - \ell(i-1)) \right] \right] && (15) \\
&\leq \mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} \left[\frac{\Phi(i, R+1+c) + \Phi(i, R-1-c)}{2} \right] && (16) \\
&+ \mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} \left[\mathbf{E}_{y \sim Q(i-1)(R)} \left[(y - \ell(i-1)) \frac{\Phi(i, R+1+c) - \Phi(i, R-1-c)}{2} \right] \right] && (17)
\end{aligned}$$

The final step is to show that the term (17) is equal to zero. As $\ell(i-1)$ is a constant with respect to R and y the term (17) can be written as:

$$\begin{aligned}
& \mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} \left[\mathbf{E}_{y \sim Q(i-1)(R)} \left[(y - \ell(i-1)) \frac{\Phi(i+1, R+1) - \Phi(i+1, R-1)}{2} \right] \right] && (18) \\
= \; & \mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} \left[B(i-1, R) \frac{\Phi(i, R+1+c) - \Phi(i, R-1-c)}{2} \right] && (19) \\
- \; & \ell(i) \mathbf{E}_{R \sim \mathbf{\Psi}(i-1)} \left[\frac{\Phi(i, R+1+c) - \Phi(i, R-1-c)}{2} \right] && (20) \\
= \; & 0 && (21)
\end{aligned}$$

□

We find that the lower bound corredponds to an unbiased random walk with step size ± 1. The upper bound also corresponds to a an unbiasd random walk with step size $\pm(1+c)$. The natural setting in the natural game is $c = 1$, which means that there is a significant difference between the upper and lower bounds. As we show in the next section, this gap converges to zero in the continuous time setting, and the upper and lower bounds match, making the strategies for both sides min-max optimal.

Note also that the adversarial strategy the aggregate loss $\ell(t)$ is always zero, regardless of the strategy of the learner, and state progression is independent of the learner's choices.

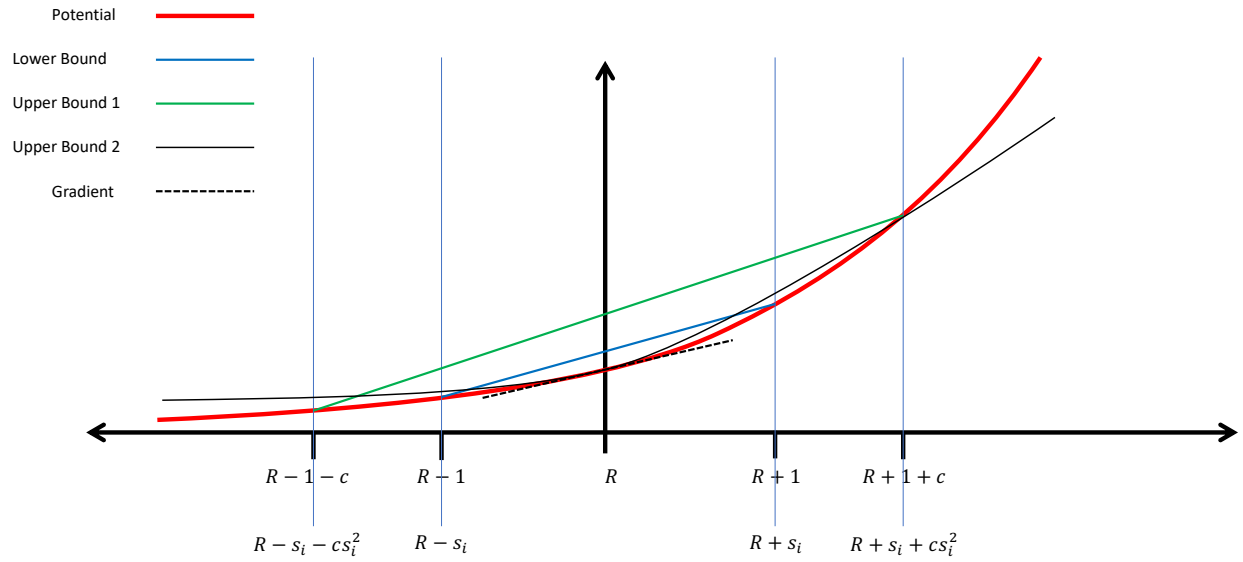

Figure 1: This figure depicts the relationship between the different upper and lower bounds used in the analysis. To aid understanding we describe the elements of the figure twice: once for the integer game, and once for the continuous time game.
Integer time game: let the current iteration be i and let the current regret be R. Let r be the regret at iteration $i+1$, we have that $R-1-c \leq r \leq R+1+c$. The potential at iteration $i+1$ is $\Phi(i+1,r)$ (the red curve). The lower bound (blue line) corresponds to the adversarial strategy: $Q^{1/2}(i,R)$. The first-order learner strategy: $P^1(i,R)$ corresponds to the green line. The second-order learner strategy: $P^2(i,R)$ corresponds to the black curve.
Continuous time game: let the current iteration be i, the current time be t_i and the current regret be R. Let $0 < s_i \leq 1$ be the step size chosen by the adversary, so that the next time is $t_{i+1} = t_i + s_i^2$. Let r be the regret at iteration $i+1$, we have that $R - s_i - cs_i^2 \leq r \leq R + s_i + cs_i^2$. The potential at iteration $i+1$ is $\Phi(t_{i+1}, r)$ (the red curve). The lower bound (blue line) corresponds to the adversarial strategy: $Q^{1/2}_{\pm s_i}(t_i, R)$. The first-order learner strategy: $P^{1c}(t_i, R)$ corresponds to the green line. The second-order learner strategy: $P^{2c}(i,R)$ corresponds to the black curve. Observe that when $s_i \to 0$ the ratio $\frac{s_i}{s_i+cs_i^2}$ converges to 1, and the upper and gap between the green and blue lines converges to zero.

3.2 A learner strategy with a variance-dependent bound

As shown in Lemma 1, the adversary always prefers mixed strategies that assign zero probability for all steps other than ± 1. Suppose, however, that the adversary is not worst-case optimal and chooses steps whose length is less than one. The following lemma gives a slightly different strategy for the learner, which guarantees a tighter bound for this case.

Lemma 3. *The learner strategy:*

$$P^2(i-1,R) = \frac{1}{Z}\frac{\partial}{\partial r}\bigg|_{r=R} \Phi(i,r) \tag{22}$$

Where Z is a normalization factor

$$Z = \mathbf{E}_{R\sim\Psi(i)}\left[\frac{\partial}{\partial r}\bigg|_{r=R} \Phi(i,r)\right]$$

guarantees the following upper potential against any adversarial strategy Q

$$\Phi^{\downarrow}(i-1,R) = \Phi(i,R) + b(i,R)\mathbf{E}_{l\sim Q(i,R)}\left[l^2\right] \tag{23}$$

where $b(i,R) = \Phi(i,R+1+c) - \Phi(i,R) - (1+c)\,\frac{\partial}{\partial r}\big|_{r=R}\,\Phi(i,r)$

We compare the bound for P^2 to the bound for P^1 given in Lemma 2. We find that when the adversary is optimal: $\mathbf{E}_{l\sim Q(t,R)}\left[l^2\right] = 1$ then the bound for P^1 is better than the bound for P^2, on the other hand, when $\mathbf{E}_{l\sim Q(t,R)}\left[l^2\right]$ is close to zero, P^2 is better than P^1.

4 Continuous time game

We start with motivation for using time that is indexed by real values rather than the natural numbers. We distinguish between two notions of time. The first notion of time is a counter that counts the iterations of the game, we will call this counter the *iteration counter* and denote it by $i = 0, 1, \ldots$. The second, more interesting notion of time is the time that appears in the regret bounds, we denote this time by t_i where i is the iteration number. We restrict the time increments $\Delta t_i = t_i - t_{i-1}$ to the range $0 \le \Delta t_i \le 1$. The magnitude of Δt_i corresponds to the *hardness* of iteration i. $\Delta t_i = 0$ corresponds to the case where the losses of all of the actions are equal to a common value $-1 \le a \le 1$. In this case the aggregate loss $\ell = a$, the state does not change: $\mathbf{\Psi}(i-1) = \mathbf{\Psi}(i)$, $\Delta t_i = 0$ and the instantanous regret is zero. On the other hand $\Delta t_i = 1$ corresponds to the adversarial strategy $Q^{1/2}(t-1,R)$ (Eqn. 7) which maximizes the regret. By allowing Δt_i to vary from iteration to iteration we get a more refined quantification of the regret and, as we show below, min/max optimality.

To find the relationship between loss magnitude and time increments we compare two adversarial strategies. The first strategy, discussed above, generates losses ± 1 with equal probability, we deonte this strategy by $Q^{1/2}_{\pm 1}$. The other strategy, denoted $Q^{1/2}_{\pm 1/k}$, generates losses of $\pm 1/k$ with equal probabilities.

From the adversarial point of view $Q^{1/2}_{\pm 1/k}$ is worse than $Q^{1/2}_{\pm 1}$. So it should correspond to a smaller time increment. But how much smaller? Suppose we start with the initial state $\mathbf{\Psi}(0)$ which is a delta functions at $R = 0$. One iteration of $Q^{1/2}_{\pm 1}$ results in a distribution ± 1 w.p, $(1/2, 1/2)$, which has mean 0 and variance 1. Suppose we associate $\Delta t = 1/j$ with a single step of $Q^{1/2}_{\pm 1/k}$. Equivalently, we associate j iterations of $Q^{1/2}_{\pm 1/k}$ with $t = 1$. How should we set j? the distribution generated by j steps is a binomial distribution supported on $j+1$ points, so there is no hope of making the two distributions identical. However, as it turns out, it is enough to equalize the mean and the variance of the two distributions. The mean of $Q^{1/2}_{\pm 1/k}$ is zero for any k. As for the variances, a single iteration of $Q^{1/2}_{\pm 1}$ is 1 and a single iteration of $Q^{1/2}_{\pm 1/k}$ is $1/k^2$. It follows that the variance after j iterations of $Q^{1/2}_{\pm 1/k}$ j/k^2. Equating this variance with that of a single step of $Q^{1/2}_{\pm 1}$ we get $j = k^2$ and $\Delta t = 1/k^2$.

Note a curious behaviour of the *range* of R as $k \to \infty$ the number of steps increases like k^2 while the size of each step is $1/k$. This means that the range of R is $[-k, k]$, which becomes converges to $(-\infty, +\infty)$ when $k \to \infty$. On the other hand, the variance increases like t.

Next lets consider effect of reducing the step size on a *biased* strategy $Q^{1/2+\gamma}_{\pm 1}$ as defined in Eqn (5) for some $0 \le \gamma \le 1/2$. We now figure out what γ' should be so that the distribution generated by k^2 iterations of $Q^{1/2+\gamma'}_{\pm 1/k}$ has the same mean as a single iteration of $Q^{1/2+\gamma}_{\pm 1}$. The mean of a single iteration of $Q^{1/2+\gamma}_{\pm 1}$ is 2γ while the mean of a single iteration of $Q^{1/2+\gamma'}_{\pm 1/k}$ is $2\gamma'/k$. Therefor to keep the means equal we need to set $2\gamma'/k = 2\gamma$ or $\gamma' = \gamma/k$.

Note that as $k \to \infty$, $\gamma' \to 0$. This observation motivates scaling the bound on $\ell(t)$ like cs_i^2 (see the description of the game below.)

This leads to the following formulation of a continuous time game. The game is a generalization of the integer time game in that it reduces to the integer time game if the adversary always chooses $s_i = 1$.

In this game we use $i = 1, 2, 3, \ldots$ as the iteration index. We use t_i to indicate a sequence of real-valued time points. $t_0 = 0$ and we assume there exists a finite n such that $t_n = T$.

We will later give some particular potential functions for which no a-priori knowledge of the termination condition is needed. The associated bounds will hold for any iteration of the game.

On iteration $i = 1, 2, \ldots$

1. If $t_{i-1} = T$ the game terminates.
2. The adversay chooses a *step size* $0 < s_i \leq 1$, which advances time by $t_i = t_{i-1} + s_i^2$.
3. Given s_i, the learner chooses a distribution $P(i)$ over $\mathbb{R}$.
4. The adversary chooses a mapping from $\mathbb{R}$ to distributions over $[-s_i, +s_i]$: $Q(t) : \mathbb{R} \to \Delta^{[-s_i,+s_i]}$
5. The aggregate loss is calculated:
$$\ell(t_i) = \mathbf{E}_{R\sim\mathbf{\Psi}t_i}\left[P(t_i, R)B(t_i, R)\right]$$
6. the aggregate loss is restricted $|\ell(t_i)| \leq cs_i^2$.
7. The state is updated. The expectation below is over distributions. and the notation $G \oplus R$ means that distribution G over the reals is shifted by the amount defined by the scalar R:
$$\mathbf{\Psi}(t_i) = \mathbf{E}_{R\sim\mathbf{\Psi}(t_{i-1})}\left[Q(t_i)(R) \oplus (R - \ell(t_i))\right]$$

When $t_i = T$ the game is terminated, and the final value is calculated:

$$\phi(T) = \mathbf{E}_{R\sim\mathbf{\Psi}(T)}\left[\Phi_{\mathrm{T}}(R)\right]$$

4.1 The adversary prefers smaller steps

As noted before, if the adversary chooses $s_i = 1$ for all i the game reduces the the integer time game. The question is whether the adversary would prefer to stick with $s_i = 1$ or instead prefer to use $s_i < 1$. In this section we give a surprising answer to this question – the adversary always prefers a smaller value of s_i to a larger one. This leads to a preference for $s_i \to 0$, as it turns out, this limit is well defined and corrsponds to Brownian motion, also known as Wiener process.

Consider a sequence of adversarial strategies S_k indexed by $k = 0, 1, 2,$. The adversarial strategy S_k is corresponds to always choosing $s_i = 2^{-k}$, and repeating $Q^{1/2}_{\pm 2^{-k}}$ for $T2^{2k}$ iterations. This corresponds to the distribution created by a random walk with $T2^{2k}$ time steps, each step equal to $+2^{-k}$ or -2^{-k} with probabilities $1/2, 1/2$. Note that in order to preserve the variance, halving the step size requires incresing the number of iterations by a factor of four.

Let $\Phi(S_k, t, R)$ be the value associated with adversarial strategy S_k, time t (divisible by 2^{-2k}) and location R. We are ready to state our main theorem.

Theorem 4. *If the final value function has a strictly positive fourth derivative:*

$$\frac{d^4}{dR^4}\Phi_T(R) > 0, \forall R$$

then for any integer $k > 0$ and any $0 \leq t \leq T$, such that t is divisible by 2^{-2k} and any R,

$$\Phi(S_{k+1}, t, R)) > \Phi(S_k, t, R)$$

Before proving the theorem, we describe it's consequence for the online learning problem. We can restrict Theorem 4 for the case $t = 0$,$R = 0$ in which case we get an increasing sequence:

$$\Phi(S_1, 0, 0) < \Phi(S_2, 0, 0) < \cdots < \Phi(S_k, 0, 0) <$$

The limit of the strategies S_k as $k \to \infty$ is the well studied Brownian or Wiener process. The backwards recursion that defines the value function is the celebrated Backwrds Kolmogorov Equation with zero dift and unit variance

$$\frac{\partial}{\partial t}\Phi(t, R) + \frac{1}{2}\frac{\partial^2}{\partial R^2}\Phi(t, R) = 0 \tag{24}$$

Given a final value function with a strictly positive fourth derivative we can use Equation (24) to compute the value function for all $0 \leq t \leq T$. We will do so in he next section.

We now go back to proving Theorem 4. The core of the proof is a lemma which compares, essentially, the value recursion when taking one step of size 1 to four steps of size 1/2.

Consider the advesarial strategies S_k and S_{k+1} at a particular time point $0 \leq t \leq T$ such that t is divisible by $\Delta t = 2^{-2k}$ and at a particular location R. Let $t' = t + \Delta t$, and fix a value function for time , $\Phi(t', R)$ and compare between two values at R, t. The first value denoted $\Phi_k(t, R)$ corresponds to S_k, and consists of a single random step of $\pm 2^{-k}$. The other value $\Phi_{k+1}(t, R)$ corresponds to S_{k+1} and consists of four random steps of size $\pm 1/2$.

Lemma 5. *If $\Phi(t', R)$ is, as a function of R continuous, strictly convex and with a strictly positive fourth derivative. Then*

- $\Phi_k(t, R) < \Phi_{k+1}(t, R)$
- *Both $\Phi_k(t, R)$ and $\Phi_{k+1}(t, R)$ are continuous, strictly convex and with a strictly positive fourth derivative.*

Proof. Recall the notations $\Delta t = 2^{-2k}$ $t' = t + \Delta t$ and $s = 2^{-k}$. We can write out explicit expressions for the two values:

- For strategy S_0 the value is

$$\Phi_k(t, R) = \frac{\Phi(t', R + s) + \Phi(t', R - s)}{2}$$

  .

- For strategy S_1 the value is

$$\Phi_{k+1}(t, R) = \frac{1}{16}(\Phi(t', R + 2s) + 4\Phi(t', R + s) + 6\Phi(t', R) + 4\Phi(t', R - s) + \Phi(t', R - 2s))$$

  .

We want to show that $\Phi_1(T - 1, R) > \Phi_0(T - 1, R)$ for all R, in other words we want to characterize the properties of Φ_{T} the would garantee that

$$\Phi_1(t, R) - \Phi_0(t, R) = \frac{1}{16}(\Phi(t', R + 2) - 4\Phi(t', R + 1) + 6\Phi(t', R) - 4\Phi(t', R - 1) + \Phi(t', R - 2)) > 0 \tag{25}$$

Inequalities of this form have been studied extensively under the name "divided differences" [4, 2, 3]. A function Φ_{T} that satisfies inequality 25 is said to be *4'th order convex* (see details in in [2]).

n-convex functions have a very simple characterization:

Theorem 6. *Let f be a function with is differentiable up to order n, and let $f^{(n)}$ denote the n'th derivative, then f is n-convex (n-strictly convex) if and only if $f^{(n)} \geq 0$ ($f^{(n)} > 0$).*

We conclude that if $\Phi(t', R)$ has a strictly positive fourth derivative then $\Phi_{k+1}(t, R) > \Phi_k(t, R)$ for all R, proving the first part of the lemma.

The second part of the lemma follows from the fact that both $\Phi_{k+1}(t, R)$ and $\Phi_k(t, R)$ are convex combinations of $\Phi(t, R)$ and therefor retain their continuity and convexity properties.

□

Proof. of Theorem 4
The proof is by double induction over k and over t. For a fixed k we take a finite backward induction over $t = T - 2^{-2k}, T - 2 \times 2^{-2k}, T - 3 \times 2^{-2k}, \cdots, 0$. Our inductive claims are that $\Phi_{k+1}(t, R) > \Phi_k(t, R)$ and $\Phi_{k+1}(t, R)$,$\Phi_k(t, R)$ are continuous, strongly convex and have a strongly positive fourth derivative. That these claims carry over from $t = T - i \times 2^{-2k}$ to $t = T - (i + 1) \times 2^{-2k}$ follows directly from Lemma 5.

The theorem follows by forward induction on k.

□

5 Strategies for the Learner in the continuous time game

The strategies we propose for the learner in the continuous time game are an adaptation of the strategies P^1, P^2 to the case where $s_i < 1$.

We start with the high-level idea. Consider iteration i of the continuous time game. We know that the adversary prefers s_i to be as small as possible. On the other hand, the adversary has to choose some $s_i > 0$. This means that the adversary always plays sub-optimally. Based on s_i the learner makes a choice and the adversary makes a choice. As a result the current state $\mathbf{\Psi}(t_{i-1})$ is transformed to $\mathbf{\Psi}(t_i)$. To choose it's action, the learner needs to assign value possible states $\mathbf{\Psi}(t_i)$. How can she do that? By assuming that in the future the adversary will play optimally, i.e. setting s_i arbitrarily small. While the adversary cannot be optimal, it can get arbitrarily close to optimal, which is brownian motion.

Solving the backwards Kolmogorov equation with the boundary condition $\Phi(T, R)$ yields $\Phi(t, R)$ for any $R \in \mathbb{R}$ and $t \in [0, T]$. We now explain how using this potential function we derive strategies for the the learner.

Note that the learner chooses a distribution *after* the adversary set the value of s_i. The continuous time version of P^1 (Eqn 9) is

$$P^{1c}(t_{i-1}, R) = \frac{1}{Z} \frac{\Phi(t_i, R + s_{i-1} + cs_{i-1}^2) - \Phi(t_i, R - s_{i-1} - cs_{i-1}^2)}{2} \tag{26}$$

Next, we consider the continuous time version of P^2 (Eqn 22)

$$P^{2c}(t_{i-1}, R) = \frac{1}{Z} \left. \frac{\partial}{\partial r} \right|_{r=R} \Phi(t_{i-1} + s_{i-1}^2, r) \tag{27}$$

6 Two self-consistent value functions

The value functions, $\Phi(t, R)$ is a solution of PDE (24):

$$\frac{\partial}{\partial t}\Phi(t, R) + \frac{1}{2}\frac{\partial^2}{\partial R^2}\Phi(t, R) = 0 \tag{28}$$

under a boundary condition $\Phi(T, R) = \Phi_{\mathrm{T}}(R)$, which we assume is continuous, conveRand has a strictly positive fourth derivative.

So far, we assumed that the game horizon T is known in advance. We now show two value functions where knowledge of the horizon is not required. Specifically, we call a value function $\Phi(t, R)$ *self consistent* if it is defined for all $t > 0$ and if for any $0 < t < T$, setting $\phi(T, R)$ as the final potential and solving for the Kolmogorov Backward Equation yields $\phi(t, R)$ regarless of the time horizon T.

We consider two solutions to the PDE:

- The exponential value function, which corresponds to exponential weights algorithm:

$$\Phi_{\exp}(R,t) = e^{\sqrt{2}\eta R - \eta^2 t}$$

  Where $\eta > 0$ is the learning rate parameter.

- The NormalHedge value:

$$\Phi_{\mathrm{NH}}(R,t) = \begin{cases} \frac{1}{\sqrt{t+\epsilon}} \exp\left(\frac{R^2}{2(t+\epsilon)}\right) & \text{if } R \geq 0 \\ \frac{1}{\sqrt{t+\epsilon}} & \text{if } R < 0 \end{cases} \tag{29}$$

  Where $\epsilon > 0$ is a small constant.

7 NormalHedge yields the fastest increasing potential

Up to this point, we considered any continuous value function with strictly positive derivatives 1-4. We characterized the min-max strategies for any such function. It is time to ask whether value functions can be compared and whether there is a "best" value function. In this section we give an informal argument that NormalHedge is the best function. We hope this argument can be formalized.

We make two observations. First, the min-max strategy for the adversary does not depend on the potential function! (as long as it has strictly positive derivatives). That strategy corresponds to the brownian process.

Second, the argument used to show that the regret relative to ϵ-fraction of the expert is based on two arguments

- The average value function does not increase with time.
- The (final) value function increases rapidly as a function of R

The first item is true by construction. The second argument suggests the following partial order on value functions. Let $\Phi_1(t,R), \Phi_2(t,R)$ be two value functions such that

$$\lim_{R\to\infty} \frac{\Phi_1(t,R)}{\Phi_2(t,R)} = \infty$$

then Φ_1 *dominates* Φ_2, which we denote by, $\Phi_1 > \Phi_2$.

On the other hand, if the value function increases too quickly, then, when playing against brownian motion, the average value will increase without bound. Recall that the distribution of the brownian process at time t is the standard normal with mean 0 and variance t. The question becomes what is the fastest the value function can grow, as a function of R and still have a finite expected value with respect to the normal distribution.

The answer seems to be NormalHedge (Eqn. 29). More precisely, if $\epsilon > 0$, the mean value is finite, but if $\epsilon = 0$ the mean value becomes infinite.

References

[1] Richard Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

[2] Saad Ihsan Butt, Josip Pečarić, and Ana Vukelić. Generalization of popoviciu-type inequalities via fink's identity. *Mediterranean journal of mathematics*, 13(4):1495–1511, 2016.

[3] Carl de Boor. Divided differences. *arXiv preprint math/0502036*, 2005.

[4] Tiberiu Popoviciu. Sur certaines inégalités qui caractérisent les fonctions convexes. *Analele Stiintifice Univ."Al. I. Cuza", Iasi, Sectia Mat*, 11:155–164, 1965.

[5] Robert E Schapire. Drifting games. *Machine Learning*, 43(3):265–291, 2001.