Continuous drifting games

Yoav Freund
AT&T Labs – Research
Shannon Laboratory
180 Park Avenue, Room A205
Florham Park, NJ 07932, USA

Manfred Opper
Department of Applied Sciences and Engineering
Aston University
B4 7ET Birmingham UK

January 25, 2002

Abstract

We combine the results of [13] and [8] and derive a continuous variant of a large class of drifting games. Our analysis furthers the understanding of the relationship between boosting, drifting games and Brownian motion and yields a differential equation that describes the core of the problem.

1 Introduction

In [7], Freund shows that boosting is closely related to a two party game called the "majority vote game". In the last year this work was extended in two ways.

First, in [13] Schapire generalizes the majority vote game to a much more general set of games, called "drifting games". He gives a recursive formula for solving these games and derives several generalizations of the boost-by-majority algorithm. Solving the game in this case requires numerical calculation of the recursive formula.

Second, in [8], Freund derives an adaptive version of the boost-by-majority algorithm. To do that he considers the limit of the majority vote game when the number of boosting rounds is increased to infinity while the advantage of each vote over random guessing decreases to zero. Freund derives the differential equations that correspond to this limit and shows that they are closely related to the equations that describe the time evolution of the density of particles undergoing Brownian motion with drift.

In this paper we combine the results of [13] and [8] and show, for a large set of drifting games, that the limit of small steps exists and corresponds to a type of Brownian motion. This limit yields a non-linear differential equation whose solution gives the min-max strategy for the two sides of the game.

We derive the analytical solution of the differential equations for several one-dimensional problems, one of which was previously solved numerically by Schapire in [13].

Our results show that there is a deep mathematical connection between Brownian motion, boosting and drifting games. This connection is interesting in and of itself and might have applications elsewhere. Also, by using this connection we might be able to derive adaptive boosting algorithms for other problems of interest, such as classification into more than two classes and regression.

The paper is organized as follows. In Section 2 we give a short review of drifting games and the use of potential functions in their analysis. In the same section we motivate the study of *continuous* drifting games. In Section 3 we restrict our attention to drifting games in which the set of allowed steps is finite and obeys conditions that we call "normality" and "regularity". We show that the recursive equation for normal drifting games, when the drift parameter δ is sufficiently small, have a particularly simple form. In Section 5 we show why it makes sense to scale the different parameters of the drifting game in a particular way when taking the small-step limit. In Section 6 we take this limit and derive the differential equations that govern the game in this limit. In Section 7 we give a physical interpretation of these differential equations and the game. We conclude with some explicit solutions in Section 8.

2 Background

2.1 The drifting game

This game was first described by Schapire in [13], here we present the game using slightly altered notation and terminology. This notation fits better with the extension of the game to the continuous domain developed in this paper.

The drifting game is a game between two opponents: "shepherd" and an "adversary". The shepherd is trying to get m sheep into a desired area, but has only limited control over them. The adversary's goal is to keep as many of the sheep as possible outside the desired area. The game consists of T rounds, indicated by $t = 1, \ldots, T$.

The definition of a drifting game consists of the following things:

- Z an inner-product vector space over which the norm $\|\cdot\|_p$ is defined.
- B a subset of Z which defines the steps the sheep can take.
- $L : Z \to R$ a loss function that associates a loss with each location.
- $\delta > 0$ is a real valued parameter which indicates the average drift required of the adversary. Larger values of δ indicate a stronger constraint on the adversary. In this paper we study the game for very small values of δ.

The game proceeds as follows. Initially, all the sheep are in the origin, which is indicated by $\mathbf{s}_i^0 = \mathbf{0}$ for all $i = 1, \ldots, m$. Round t consists of the following steps:

1. The shepherd chooses weight vectors $\mathbf{w}_i^t$ for each sheep $i = 1, \ldots, m$.
2. The adversary chooses a step vector for each sheep $\zeta_i^t \in B$: such that
$$\sum_{i=1}^{m} \mathbf{w}_i^t \cdot \zeta_i^t \geq \delta \sum_{i=1}^{m} \left\| \mathbf{w}_i^t \right\|_p \tag{1}$$
3. The sheep move: $\mathbf{s}_i^{t+1} = \mathbf{s}_i^t + \zeta_i^t$.

After the game ends, and the position of the sheep are $\mathbf{s}_i^{T+1}$, the shepherd suffers the final average loss:

$$L = \frac{1}{m} \sum_{i=1}^{m} L(\mathbf{s}_i^{T+1})$$

.

2.2 Why are drifting games interesting?

Drifting games were introduced by Schapire in [13] as an abstraction which generalizes the boost-by-majority algorithm [7] and the binomial weights learning algorithm [5]. In this subsection we describe these relations and provide the motivation for studying drifting games. In the next subsection we motivate the study of *continuous drifting games*.

The boost-by-majority algorithm corresponds to a very simple drifting game, the adversary controls the weak learning algorithm, and the shepherd corresponds to the boosting algorithm, the other elements of the game are defined as follows

- Z is the real line and the norm $\|\cdot\|_p$ is the absolute value.
- B is the set $\{-1, +1\}$.
- $L : Z \to R$ is defined to be 0 for s ≤ 0 and 1 otherwise.
- $\delta > 0$ corresponds to the requirement that the error of the generated weak hypotheses be at most $(1 - \delta)/2$.

In this game each sheep corresponds to a training example. The location of the sheep on the real line corresponds to the number of hypotheses whose prediction on the example is correct. Each hypothesis corresponds to moving those examples on which the hypothesis is correct one step down and the rest of the examples one step up. The loss function L associates a loss of 1 with those examples on which the majority vote over the weak hypotheses is incorrect. In this game the vectors chosen by the shepherd correspond to the weights the boosting algorithm places on the different examples. For complete details of this correspondence see Freund [7].

Interestingly, this very same game, with a different interpretation for sheep and locations, corresponds to the problem of online learning with expert advice as well as to an interesting variant of the twenty-one questions game. The online learning problem was studied by Cesa-Binachi et. al in [5]. In this case the sheep correspond to the experts and the location of the sheep corresponds to the number of mistakes made by the experts. An assumption is made that there is an expert which makes no more than k mistakes, the identity of this expert is unknown. The goal is to design an algorithm for combining the advice of the experts which is guaranteed to make the least number of mistakes under the stated assumption. The analysis of this game yields the binomial weights algorithm.

The exact same game was described by Spencer in [15] who called it the "balancing vector" game or the "pusher-chooser" game. Aslam and Dhagat refer to the same game as a "chip" game and used it to analyze a variety of problems (see [2, 6, 3, 1]).

One application of this game which is closely related to the online learning problem is the twenty questions game, also called "Ulam's game", with a fixed number of lies. This game was studied by Dhagat et al. [6] and by Spencer [16]. In the twenty questions game one party holds a secret represented by an integer number in some range $1, \ldots, N$. The second party tries to identify the secret number by asking questions of the form: "Is the number in the set X?". Clearly, the optimal strategy for this game is to perform a binary search. The game becomes much more interesting if the party holding the secret is allowed to lie up to k times. Where k is an a-priori fixed parameter of the game. The analysis of this game by Dhagat et. al. and by Spencer is based on the same drifting game (here called a chip game) as both of the previous problems. In this case each sheep corresponds to a possible secret value and the location of the sheep corresponds to the number of lies that have been made assuming this was the chosen secret. The solution given to this game is essentially identical to the binomial weights algorithm.

Generalizing the drifting game from this simple one-dimensional case provides a general method for designing algorithms for a large class of problems. Among them boosting algorithms for learning problems in which there are more than two possible labels.

Consider first the case where the number of possible labels is $k > 2$. In [13] Schapire describes a reduction of this problem to a drifting game on a R^{k-1}. This reduction suggests a boosting algorithm for the non-binary learning setup, however, a closed form solution to this game is not known and the calculation of the solution is computationally expensive (on the other hand, it can be done prior to receiving the training data). At the end of this paper we show how the techniques we develop here can be used to calculate the optimal strategy for this drifting game for the case $k = 3$.

Next, consider designing a boosting algorithm for learning problems in which the label is a real valued number. An insightful way of looking at this problem is to think of boosting as a variational optimization method as was suggested by Mason et al. [11, 12]. In this view boosting is seen as a gradient descent method in function space. Each boosting iteration corresponds to adding a simple function to the existing solution with the goal of minimizing the average loss over the training examples. Mason et al. choose to use the *same* loss function as the guide for all the boosting iterations. Intuitively, this choice is not optimal because the effect of a local improvement in the approximation in the first boosting iteration is significantly smaller than the effect of a similar improvements at the last iteration. This is because early improvements are in danger of being corrupted by subsequent iterations while improvements done on the last iteration are guaranteed to stay as they are. The drifting game is a natural model for this effect. By using the drifting game analysis we can design boosting algorithms for arbitrary loss functions. The meaning of "weak hypotheses" in this case are hypotheses whose value is has a slight negative correlation with the error of the approximation function. In the last section of this paper we show how to calculate the optimal time-dependent loss function for the target losses $(y - y')^2$ and $\min(1, (y - y')^2)$.

2.3 Why is the continuous limit interesting?

In this paper we consider limits of drifting games in which the size of the steps decreases to zero, the number of steps increases to infinity and δ decreases to zero. The exact form of the limit will be justified in Section 5. In this section we motivate exploration of this limit.

Our main interest in this limit stems from the fact that it gives us a general way of designing boosting algorithms that are *adaptive*. The Adaboost algorithm [9, 14] was the first adaptive boosting algorithm and this adaptivity is one of the main reasons that Adaboost had a much more significant impact on practical machine learning than its two predecessors. We say that Adaboost is adaptive because there is no need to make an a-priori assumption on the maximal error of the weak hypotheses that will be generated during boosting. Instead, the algorithm adapts to the accuracy of the hypothesis that it received from the weak learner, stronger hypotheses cause larger changes in the weights assigned to the examples and weaker hypotheses cause smaller changes. In [8] Freund presented the Brownboost algorithm which is an adaptive variant of the boost by majority algorithm. The transition from boost-by-majority to BrownBoost is done by taking the continuous time limit. Freund also shows that in that paper that Adaboost is a special case of the resulting boosting algorithm.

In this paper we combine Freund's Brownboost algorithm with Schapire's drifting games [13] and in this way show how to make any boosting algorithm (which can be represented as a drifting game) adaptive.

Another benefit of taking the continuous limit is simple and elegant mathematical structure of the resulting game. While Schapire's solution of the drifting games is very general, its calculation requires the solution of a complex recursion involving minima and maxima. On the other hand, as we shall show in section 4, when δ is sufficiently small, the recursion becomes simpler and does not involve minima and maxima. Taking the limit of this simplified recursion yields a stochastic differential equation that characterizes the optimal potential function for the game.

The stochastic differential equation is identical to the equations that characterize Brownian motion. These equations have been studied extensively both in physics, in statistical mechanics, and in mathematical economics, in the study of option pricing and the Black-Scholes Equation.[1] Using this existing work, we can use techniques that have been developed for solving stochastic differential equations to calculate the optimal strategies for drifting games. In some cases we can do find closed form analytical representation of the optimal potential function. In other cases, when finding a closed form solution is too hard, it is often possible to use existing numerical algorithms for solving partial differential equations to find a numerical solution.[2] In Section 8 we provide analytical and numerical solutions to some simple drifting games.

Finally, the connection established here between Brownian motion and drifting games might provide new insights into the workings of systems in which a weak global force field acts on a large number of small elements. Such systems seem to be quite common in physics and in economics.

2.4 Analysis of drifting games using a potential function

In [13] Schapire shows that drifting games can be solved by defining a potential function $\phi_t(\mathbf{s})$. Setting the boundary condition $\phi_T(\mathbf{s}) = L(\mathbf{s})$ and solving the recursion:

$$\phi_{t-1}(\mathbf{s}) = \min_{\mathbf{w}} \sup_{\mathbf{z}} \left\{ \phi_t(\mathbf{s}+\mathbf{z}) + \mathbf{w} \cdot \mathbf{z} - \delta \left\| \mathbf{w} \right\|_p \right\} \quad (2)$$

The minimizing vector $\mathbf{w}$ defined the weight vectors that are the min/max strategy for the shepherd.

One can show (Theorem 1 in [13]) that the average potential is non increasing

[1] For a wonderful review of the mathematical underpinning of the Black Scholes equation and a new game-theoretic analysis of option pricing, see the recent book by Shafer and Vovk [**?**].

[2] Most of these algorithms use finite element methods that are very similar to the simplified recursions described above. However, the advantage of using an existing debugged and optimized software package rather than writing one by yourself should not be overlooked!

$$\sum_i \phi_t(\mathbf{s}_i^{t+1}) \leq \sum_i \phi_{t-1}(\mathbf{s}_i^t)$$

Hence, one gets the bound on the average loss

$$\frac{1}{m}\sum_{i=1}^{m} L(\mathbf{s}_i^{T+1}) \leq \phi_0(\mathbf{s} = \mathbf{0}) \tag{3}$$

3 Normal and regular lattices

We assume that the sheep positions $\mathbf{s}_i^t$ are vectors in R^d. We restrict the set of allowed steps B to be a finite set of size $d+1$ $\mathbf{z}_0, \ldots, \mathbf{z}_d$ which spans the space R^d and such that $\sum_{i=0}^{d} \mathbf{z}_i = 0$. If a set B satisfies these conditions, we say it is *normal*.

If the set B is normal and, in addition, satisfies the following two symmetries for some positive constants a and b, we say it is *regular*.

1. For any $i \in 0, \ldots, d$, $\mathbf{z}_i \cdot \mathbf{z}_i = a$
2. For any $i, j \in 0, \ldots, d$ such that $i \neq j$, $\mathbf{z}_i \cdot \mathbf{z}_j = -b$

For example, a regular set in R^2 is

$$\mathbf{z}_1 = (0, 1),\ \mathbf{z}_2 = \frac{1}{2}(\sqrt{3}, -1),\ \mathbf{z}_3 = \frac{1}{2}(-\sqrt{3}, -1) \tag{4}$$

Given an inner-product vector space whose dimension is at least d it is easy to construct a regular set B of size $d+1$ for this space. For any orthonormal set of size d, $\mathbf{v}_1, \ldots, \mathbf{v}_d$ we can derive a regular set by setting $\mathbf{z}_0, \ldots, \mathbf{z}_d$ to be

$$\mathbf{z}_0 = -\frac{1}{\sqrt{d}}\sum_{j=1}^{d} \mathbf{v}_j;$$

$$\forall i = 1, \ldots, d,\ \mathbf{z}_i = \sqrt{\frac{d+1}{d}}\mathbf{v}_i - \frac{\sqrt{d+1}-1}{d^{3/2}}\sum_{j=1}^{d} \mathbf{v}_j$$

4 The drifting game for small values of δ

Given that the set B is normal, we can show that, for sufficiently small values of δ, the solution of the game has a particularly simple form.

Theorem 1 *Let B be a normal set of steps. Then there exists some $\delta_0 > 0$ such that for any potential function ϕ_t and location $\mathbf{s}$ and any $\delta_0 \geq \delta \geq 0$ the solution to the recursive definition of ϕ_{t-1} satisfies*

$$\phi_{t-1}(\mathbf{s}) = \frac{\sum_{i=0}^{d} \phi_t(\mathbf{s} + \mathbf{z}_i)}{d+1} - \delta \left\| \mathbf{w}^* \right\|_p \tag{5}$$

and $\mathbf{w}^*$ is the local slope of $\phi_t(\mathbf{s})$, i.e.

$$\phi_t(\mathbf{s} + \mathbf{z}_i) = C + \mathbf{w}^* \cdot \mathbf{z}_i \ ; \quad C \doteq \frac{\sum_{j=0}^{d} \phi_t(\mathbf{s} + \mathbf{z}_j)}{d+1} \tag{6}$$

If, in addition, the set B is regular, then one can set

$$\delta_0 = \frac{1}{d} \min_{\mathbf{w} \neq 0} \frac{\|\mathbf{w}\|_2}{\|\mathbf{w}\|_p}$$

Before we prove this theorem, it is interesting to consider its implications on the (close to) optimal strategies for the two opponents in the drifting game. What we have is that, for sufficiently small values of δ, the optimal strategy for the shepherd is to set the weight vector $\mathbf{w}_i^t$ for sheep i at round t to be the *slope* of the potential function for round $t+1$ as defined for the $d+1$ locations reachable at round $t+1$ by sheep i.

Next consider the adversary, we apply the adversarial strategy described by Schapire in the proof of Theorem 2 in [13] to our case. Consider the case where the number of sheep m is very large (alternatively, one can consider "infinitely divisible" sheep.) In this case an almost-optimal strategy for the adversary is to select the step ζ_i^t of sheep i independently at random with a distribution $p_{i,j}^t$ over the $d+1$ possible steps $\mathbf{z}_0, \ldots, \mathbf{z}_d$ such that for all sheep i

$$\sum_{j=0}^{d} p_{i,j}^t \mathbf{z}_j = (\delta + \mu) \mathbf{w}_i^t \frac{\|\mathbf{w}_i^t\|_p}{\|\mathbf{w}_i^t\|_2^2}$$

For some small $\mu > 0$.

It follows that the expected value of the required average drift is

$$\mathbf{E}\left[\sum_{i=1}^{m} \mathbf{w}_i^t \cdot \zeta_i^t\right] = (\delta+\mu)\sum_{i=1}^{m} \mathbf{w}_i^t \cdot \mathbf{w}_i^t \frac{\|\mathbf{w}_i^t\|_p}{\|\mathbf{w}_i^t\|_2^2} = (\delta+\mu)\sum_{i=1}^{m} \left\|\mathbf{w}_i^t\right\|_p \tag{7}$$

As m is large and the steps are chosen independently at random the actual value of $\sum_{i=1}^{m} \zeta_i^t \cdot \mathbf{w}_i^t$ is likely to be very close to its expected value and thus, with high probability
$\sum_{i=1}^{m} \zeta_i^t \cdot \mathbf{w}_i^t > \delta \sum_{i=1}^{m} \|\mathbf{w}_i^t\|_p$. As $m \to \infty$ we can let $\mu \to 0$ and so, in the limit of very many sheep, the strategy satisfies the drifting requirement exactly.

Assuming that the adversary uses this strategy with $\mu = 0$ yields an interesting new interpretation of the potential function. It is not hard to see that $\phi_t(\mathbf{s})$ is the expected final loss of a sheep conditioned on the fact that it is located at $\mathbf{s}$ at round t.The recursive relation between the potential in consecutive rounds is simply a relation between these conditional expectations.

We now prove the theorem.

Proof:
We fix a location $\mathbf{s}$ and consider the recursive definition of $\phi_{t-1}(\mathbf{s})$.

Consider first the case $\delta = 0$. In this case the min max formula (2) can be written as

$$\begin{aligned} \phi_{t-1}(\mathbf{s}) &= \min_{\mathbf{w}} F(\mathbf{w}) \\ F(\mathbf{w}) &= \max_{i=0,\ldots,d} f_i(\mathbf{w});\ f_i(\mathbf{w}) = \phi_t(\mathbf{s}+\mathbf{z}_i) + \mathbf{w}\cdot\mathbf{z}_i \end{aligned} \tag{8}$$

Note that for each i, $f_i(\mathbf{w})$ is a simple affine function whose slope is $\mathbf{z}_i$. Thus $F(\mathbf{w})$ is a convex function whose minimum is achieved on a convex set. We shall now show that this set consists of a single point.

To test whether a point $\mathbf{w}$ is a local minimum we consider the restriction of the function $F(\mathbf{w})$ to rays emanating from $\mathbf{w}$. Given a point $\mathbf{w}$ and a direction vector $\mathbf{v}$ such that $\|\mathbf{v}\|_p = 1$, we define the function $g_{\mathbf{w},\mathbf{v}} : [0,\infty) \to (-\infty,+\infty)$ as $g_{\mathbf{w},\mathbf{v}}(x) = F(\mathbf{w}+x\mathbf{v}) - F(\mathbf{w})$.

Let $\mathbf{w}^*$ be a point on which the minimum of $F(\mathbf{w})$ is achieved and let $\mathbf{v}$ be an arbitrary direction. It is easy to verify that $g_{\mathbf{w},\mathbf{v}}(x) = x \max_i \mathbf{z}_i \cdot \mathbf{v}$. Thus $g_{\mathbf{w},\mathbf{v}}$ is constant if and only if $\max_i \mathbf{z}_i \cdot \mathbf{v} = 0$. Written in another way, this means that $\mathbf{z}_i \cdot \mathbf{v} \le 0$ for all $i = 0,\ldots,d$. Consider the two possibilities. If $\mathbf{z}_i \cdot \mathbf{v} = 0$ for all i then $\mathbf{v}$ is orthogonal to the space spanned by the $\mathbf{z}_i$'s, which contradicts the assumption that the set B is normal and thus spans the space. If there is some i for which $\mathbf{z}_i \cdot \mathbf{v} < 0$ then $\mathbf{v} \cdot (\sum_i \mathbf{z}_i) < 0$ which implies that $\sum_i \mathbf{z}_i \neq 0$ which again contradicts our assumption that B is normal. We conclude that $g_{\mathbf{w},\mathbf{v}}$ is a strictly increasing function for all $\mathbf{v}$ and thus $\mathbf{w}^*$ is the unique minimum.

The fact that the minimum is unique implies also that at the minimum all the affine functions over which we take the max are equal, $f_i(\mathbf{w}^*) = c$ for all $i = 0, \ldots, d$. Summing over i , and recalling that $\sum_i \mathbf{z}_i = 0$ we find that

$$\begin{aligned} (d+1)c = \sum_{i=0}^{d} f_i(\mathbf{w}^*) &= \sum_{i=0}^{d} (\phi_t(\mathbf{s}+\mathbf{z}_i) + \mathbf{w}^* \cdot \mathbf{z}_i) \\ &= \sum_{i=0}^{d} (\phi_t(\mathbf{s}+\mathbf{z}_i)) \end{aligned}$$

and thus the recursion yields

$$\phi_{t-1}(\mathbf{s}) = \frac{1}{d+1} \sum_{i=0}^{d} (\phi_t(\mathbf{s}+\mathbf{z}_i))$$

$$f_i(\mathbf{w}^*) = \phi_{t-1}(\mathbf{s}) \; \forall i = 0, \ldots, d$$

completing the proof of the theorem for the case $\delta = 0$.

We next consider the case $\delta > 0$. In this case we redefine $f_i(\mathbf{w})$ in Equation (8) to be

$$f_i(\mathbf{w}) = \phi_t(\mathbf{s}+\mathbf{z}_i) + \mathbf{w} \cdot \mathbf{z}_i - \delta \left\|\mathbf{w}\right\|_p \; .$$

In what follows, we will refer to the definition of F when $\delta = 0$ as F_0.

We will now show that for sufficiently small values of δ the minimizer vector $\mathbf{w}^*$ is the same as it was for $\delta = 0$. To see that, consider the directional derivative of F at a point $\mathbf{w}$ and direction $\mathbf{v}$:

$$D_{\mathbf{w},\mathbf{v}}(F) \doteq \left. \frac{dg_{\mathbf{w},\mathbf{v}}(x)}{dx} \right|_{x=0}$$

Clearly, the function $F(\mathbf{w})$ is continuous and has a directional derivative everywhere, thus a point $\mathbf{w}$ is a local minimum of $F(\mathbf{w})$ if and only if $D_{\mathbf{w},\mathbf{v}}(F) \geq 0$ for all directions $\mathbf{v}$. As the directional derivative is a linear operator, the directional derivative of $F(\mathbf{w})$ is the sum of the directional derivative of $F_0(\mathbf{w})$ and the directional derivative of $\delta \left\|\mathbf{w}\right\|_p$.

We start with F_0. As shown earlier, the ray functions for F_0 are equal to $g_{\mathbf{w},\mathbf{v}}(x) = x \max_i \mathbf{z}_i \mathbf{w}^*$. There are two cases depending on the value of $\mathbf{w}$:

- If $\mathbf{w} = \mathbf{w}^*$ then $D_{\mathbf{w}^*,\mathbf{v}}(F_0) = \max_i \mathbf{z}_i \cdot \mathbf{v} > 0$ and thus $\min_{\mathbf{v}} D_{\mathbf{w}^*,\mathbf{v}}(F) = a > 0$ where a depends only on the set B and is independent of the potential function ϕ_t.

- If $\mathbf{w} \neq \mathbf{w}^*$ then there is a line segment between $\mathbf{w}$ and $\mathbf{w}^*$ on which the function F is defined by the ray function $g_{\mathbf{w}^*,\mathbf{v}}$ where $\mathbf{v} = (\mathbf{w}^* - \mathbf{w})/\left\|\mathbf{w}^* - \mathbf{w}\right\|_p$ and thus $D_{\mathbf{w},\mathbf{v}}(F_0) = -D_{\mathbf{w}^*,-\mathbf{v}}(F_0) < a < 0$.

Consider now the directional derivative of $\delta \left\|\mathbf{w}\right\|_p$. As $\left\|\mathbf{w}\right\|_p$ is a norm, $\left\|\mathbf{w} + x\mathbf{v}\right\|_p \leq \left\|\mathbf{w}\right\|_p + x\left\|\mathbf{v}\right\|_p = \left\|\mathbf{w}\right\|_p + x$. Thus $|D_{\mathbf{w},\mathbf{v}}(\delta \left\|\mathbf{w}\right\|_p)| \leq \delta$.

Combining these two observations we find that if $\delta < a$ then

- For $\mathbf{w} = \mathbf{w}^*$, $D_{\mathbf{w}^*,\mathbf{v}}(F) > 0$ for all $\mathbf{v}$, i.e. $\mathbf{w}^*$ is a local minimum of F.

- For $\mathbf{w} \neq \mathbf{w}^*$, $D_{\mathbf{w},\mathbf{w}^*-\mathbf{w}} < 0$, i.e. $\mathbf{w}$ cannot be a local minimum of F.

We conclude that if we set $\delta_0 = a$ then for any $\delta < \delta_0$ the minimizer $\mathbf{w}^*$ is the slope of $\phi_t(\mathbf{s})$ and the formula for $\phi_{t-1}(\mathbf{s})$ is as stated in the theorem.

Finally, we identify the setting of δ_0 for a regular set B. If B is regular then we can explicitly calculate $\delta_0 = \min_{\mathbf{v}} \max_i \mathbf{z}_i \cdot \mathbf{v}$. This can be done in two steps. First, we show that the vectors $\mathbf{v}$ which achieve the minimum in the last equation are $\mathbf{v} = -\mathbf{z}_i$ for $0 \leq i \leq d$. Second, we find that if B is regular, then $\mathbf{z}_i \cdot \mathbf{z}_j / \left\|\mathbf{z}_i\right\|_2 = 1/d$ for any $i \neq j$. ■

5 Exploring different limits

Given that the solution we found for the shepherd has a natural interpretation as a type of a slope or local gradient, it is natural to consider ways in which we can generalize the game from its original form in discrete time and space to continuous time and space. Also, as was shown by Freund [8], when applying the drifting game analysis to boosting methods, it turns out that the continuous limit corresponds to the ability to make the algorithm "adaptive".

The way in which we design the continuous version of the drifting game is to consider a sequence of games, all of which use the same final loss function, in which the size of the steps become smaller and smaller while at the same time the number of steps becomes larger and larger.

Fix a loss function $L : R^d \to R$ and let B be a normal step set. We define the game G_T to be the game where the number of steps is T and the step set is

$\epsilon_T B = \{\epsilon_T \mathbf{z}_i, i = 0, \ldots, d\}$ where $\epsilon_T > 0$ and $\epsilon_T \to 0$ as $T \to \infty$.To complete the definition of the game we need to choose δ_T and ϵ_T. We do this under the assumption that δ_T is always sufficiently small so that the solution described in the previous section holds and we base our argument on the almost-optimal stochastic strategy of the adversary described there.

First, consider δ_T. If all of the drift vectors point in the same direction then the expected average location of the sheep after T steps is distance $T\delta_T$ from the origin. If $T\delta_T \to \infty$ then the average total drift of the sheep is unbounded and the shepherd can force them all to get arbitrarily far from the origin. On the other hand, if $T\delta_T \to 0$ then the shepherd loses all its influence as $T \to \infty$ and the sheep can just choose a step uniformly at random and, in the limit, reach a uniform distribution over the space. We therefore assume that $\delta_T = c_1/T$.

Next we consider ϵ_T. In this case the strategy of the adversary corresponds to simple random walk. Thus after T steps the variance of sheep distribution is $T\epsilon_T^2$. Similarly to the previous case, if $T\epsilon_T^2 \to 0$ then the adversary has too much power while if $T\epsilon_T^2 \to \infty$ the adversary is too weak. We therefore set $\epsilon^T = c_2/\sqrt{T}$.

Finally, as we let the number of rounds increase and the step size decrease, it becomes natural to define a notion of "time" τ to be t/T.

We can re-parameterize this limit by setting $\epsilon = 1/\sqrt{T}$, and absorbing c_2 into the definition of $\mathbf{z}_i$. We thus get a scaling in which $\mathbf{z}_i' = \epsilon \mathbf{z}_i$, $\delta' = \epsilon^2 \delta$ and $d\tau = \epsilon^2$. Letting $\epsilon \to 0$ we get a continuous time and space variant of the drifting game and its solution. Assuming also that the number of sheep m grows to infinity we have an optimal strategy for the adversary. This strategy, in the limit, corresponds to Brownian motion of the sheep with a location-dependent drift component.

6 The continuum limit

We will now show that the latter definition of a continuum limit also leads to a natural limit of the recursion defined in Equation (5) by a partial differential equation.

With the replacement $\mathbf{z}_i' = \epsilon \mathbf{z}_i$ and $\delta' = \epsilon^2 \delta$, assuming that ϕ can be extended to a smooth function of the continuous variable s, we expand the right hand side in a Taylor series up to second order in ϵ.

$$\phi_{\tau-1}(\mathrm{s}) - \phi_\tau(\mathrm{s}) = \epsilon \frac{1}{|B|} \sum_i \mathbf{z}_i \cdot \nabla \phi_\tau(\mathrm{s}) + \tag{9}$$

$$\frac{\epsilon^2}{2|B|}\sum_{k,l}\sum_i z_i^k z_i^l \frac{\partial^2\phi_\tau(\mathbf{s})}{\partial s^k \partial s^l} - \epsilon^2\delta\left\|\mathbf{w}\right\|_p + o(\epsilon^2)$$

We introduce the continuous time variable τ via $t\epsilon^2$, and expand the left hand side of Equation (9) to first order in ϵ^2. Finally, replacing $\phi_t(\mathbf{s})$ by $\phi(\mathbf{s},\tau)$ and dividing by ϵ^2, we get

$$\frac{\partial\phi(\mathbf{s},\tau)}{\partial\tau} = -\frac{1}{2}\sum_{k,l} D_{kl}\frac{\partial^2\phi(\mathbf{s},\tau)}{\partial s^k\partial s^l} + \delta\left\|\mathbf{w}^*\right\|_p \tag{10}$$

where

$$D_{kl} = \frac{1}{d+1}\sum_{i=0}^{d} z_i^k z_i^l$$

and z_i^k and s^k denotes the kth components ($k = 1,\ldots,d$) of the vectors $\mathbf{z}_i$ and s respectively. The linear term in ϵ vanishes due to the extra condition on the vectors $\mathbf{z}_i$. For the regular set described in Example (4), we get the diagonal matrix $D_{kl} = \frac{1}{2}$ for $k = l$ and zero otherwise. Finally, we get an explicit form for the drift vector $\mathbf{w}^*$ in the continuum limit by replacing $\mathbf{z}_i$ with $\mathbf{z}_i'$ and expanding the local slope in Equation (6) to first order in ϵ. This simply yields the gradient

$$\mathbf{w}^*(\mathbf{s},\tau) = -\nabla\phi(\mathbf{s},\tau) \tag{11}$$

Combining Equations (10) and (11) we find that the recursion for the potential in the continuum limit is given by the nonlinear partial differential equation

$$\frac{\partial\phi(\mathbf{s},\tau)}{\partial\tau} = -\frac{1}{2}\sum_{k,l} D_{kl}\frac{\partial^2\phi(\mathbf{s},t)}{\partial s^k\partial s^l} + \delta\left\|\nabla\phi(\mathbf{s},\tau)\right\|_p \tag{12}$$

7 Physical interpretation: diffusion processes

We will now come back to the probabilistic strategy of the sheep discussed in Section 3 and show that Equation (12) has a natural interpretation in the context of a *diffusion process*. Physical diffusion processes model the movement of particles in viscous media under the combined influence of a thermal random walk and a force field. The process can be described from two perspectives (see Breiman [4] for a good introduction to the mathematics of diffusion processes).

From the perspective of each single particle, the diffusion process can be seen as the continuous time limit of a random walk. From this perspective, the limit of the stochastic strategies for the sheep which is described in Section 3 is a diffusion process in which the force field is defined through the weight vectors chosen by the shepherd and the diffusion is a location independent quantity defined by the set B. Formally, a diffusion process defines a Markovian distribution over particle trajectories $\mathbf{s}(\tau)$. The trajectories are continuous but have no derivative anywhere. The distribution over trajectories is defined by the average change in the position (the drift) during a time interval h $\mathbb{E}[\mathbf{s}(\tau+h) - \mathbf{s}(t)|\mathbf{s}(t) = \mathbf{s}] = h\mathbf{A}(\mathbf{s},\tau) + o(h)$ and the variance of the change in the position (the diffusion) $\mathbb{E}\left[(s^k(\tau+h) - s^k(\tau))(s^l(\tau+h) - s^l(\tau))|\mathbf{s}(\tau) = \mathbf{s}\right] = hD_{kl} + o(h)$. Both the drift and the diffusion behave linearly for small h. [3] $\mathbf{A}$ is usually called the *drift* field and D the *diffusion* matrix. Taking the limit of small step size in Equation (7) we get

$$\mathbf{A}(\mathbf{s},\tau) = \mathbf{w}(\mathbf{s},\tau)\frac{||\mathbf{w}(\mathbf{s},\tau)||_p}{||\mathbf{w}(\mathbf{s},\tau)||_2^2} \tag{13}$$

Assuming the 2-norm, the emerging diffusion problem is that of a particle under an external force of constant modulus. The optimal strategy of the shepherd amounts in finding the direction of $\mathbf{A}$ for each position and time such that the expected loss at the final time is minimal.

The second perspective for describing a diffusion process is to consider the temporal development of the particle *density*. This development is described by the conditional density $P(\mathbf{r},\tau'|\mathbf{s},\tau)$ which describes the distribution at time τ' of a unit mass of particles located at $\mathbf{s}$ at time τ. The time evolution of this conditional distribution is described by the *forward* or *Fokker-Planck* equation:

$$\begin{aligned}\frac{\partial P(\mathbf{r},\tau'|\mathbf{s},\tau)}{\partial\tau'} = \frac{1}{2}\sum_{k,l}&\frac{\partial^2\left[P(\mathbf{r},\tau'|\mathbf{s},\tau)D_{kl}(\mathbf{r},\tau')\right]}{\partial r^k\partial r^l} \\ &- \left[\nabla_{\mathbf{r}}\cdot\mathbf{A}(\mathbf{r},\tau')\right]P(\mathbf{r},\tau'|\mathbf{s},\tau)\end{aligned} \tag{14}$$

One can show that the partial differential Equation (12) for $\phi(\mathbf{s},\tau)$ naturally comes out of this diffusion scenario by the interpretation of the potential $\phi(\mathbf{s},\tau)$ as the

[3] Note that on average the displacement (or velocity) is proportional to the force. This behavior is unlike the well known Newtonian law: "acceleration $\propto$ force" which describes the motion of free particles. In our case, the particles may be understood as moving in a highly viscous medium for which the effect of damping is much stronger than the effects of inertia.

expected loss at time $\tau' = 1$ when a sheep is at time τ at the position s i.e..

$$\phi(\mathbf{s}, \tau) = \int d\mathbf{r} L(\mathbf{r})\ P(\mathbf{r}, \tau' = 1 | \mathbf{s}, \tau) \tag{15}$$

By using the so called *Backward Equation* (see description in [10]) which describes the evolution $P(\mathbf{r}, \tau' | \mathrm{s}, \tau)$ with respect to the initial condition s and τ one arrives at Equation (12).

8 Explicit solutions for $d = 1$

In general, in order to solve partial differential equations such as Equation (12) one has to resort to numerical procedures which are based on discretization and lead to recursions similar to the ones defined in Equations (5) and (6). Nevertheless, for dimension $d = 1$ and specific classes of loss functions analytic solutions are possible.

Setting $z_{1,2} = \pm\epsilon$ and $D = 1$ Equation (12) reads

$$\frac{\partial \phi(s, \tau)}{\partial \tau} = -\frac{1}{2} \frac{\partial^2 \phi(s, \tau)}{\partial s^2} + \delta \left| \frac{\partial \phi(s, \tau)}{\partial s} \right| \tag{16}$$

Explicit solutions are possible for loss functions where time independent regions can be found for which
$w^*(s, \tau) = -\frac{\partial \phi(s,\tau)}{\partial s}$ has a constant sign. Constrained to such regions, Equation (16) is *linear*. We will discuss 2 cases next:

Monotonic loss: In this case $\frac{\partial L(s)}{\partial s} > 0$ for all $s \in R$ (alternatively $\frac{\partial L(s)}{\partial s} < 0$ for all $s \in R$.) The simplest solution is for the exponential loss $L_e(s) = e^s$. The potential function for this case is $\phi(s, \tau) = \exp(s + (1/2 + \delta)\tau)$. The weight function is $w^*(s, \tau) = (1/2 + \delta) \exp(s + (1/2 + \delta)\tau)$. Note that multiplying $w^*(s, \tau)$ by a function that depends only on τ does not change the game. Therefor an equivalent weight function is simply $w^*(s, \tau) = e^s$. This last weight function (with a reversal of the sign in the exponent) is the one used in Adaboost [9]. The exponential potential function is the one underlying all of the algorithms for online learning using multiplicative weights (see e.g. [17]). As this potential function remains unchanged (other than a constant scalar, which can usually be ignored) changing the time horizon does not change the otimal strategies. This allows us to remove the time horizon from the definition of the game. Indeed, this might be the reason that the exponential potential function is so central to the design of so many optimization algorithms.

Another known solution is for the *step* loss function: $L_{BBM}(s) = 1$ for $s < 0$ and $L_{BBM}(s) = 0$ otherwise. The solution in this case is

$$\phi(\mathbf{s}, \tau) = \frac{1}{2}\left(1 - erf\left(\frac{s + \delta(1-\tau)}{\sqrt{2(1-\tau)}}\right)\right) \quad . \tag{17}$$

This potential function is the one used in Brownboost[8].

Symmetric Loss: $L(s) = L(-s)$ leading to $\mathbf{w}^*(-s, \tau) = -\mathbf{w}^*(s, \tau)$ where $L(s)$ monotonic in $[0, \infty)$. It is often easier to solve the corresponding Fokker-Planck equation (14) setting $A(s, \tau) = \text{sign}(w^*(s, \tau))$. We illustrate this for the case of *increasing* loss functions $A(s, \tau) = 1$ for $s \geq 0$.

$$\frac{\partial P(r, \tau'|s, \tau)}{\partial \tau'} = \frac{1}{2}\frac{\partial^2 P(r, \tau'|s, \tau)}{\partial r^2} + \delta \frac{\partial P(r, \tau'|s, \tau)}{\partial r} \tag{18}$$

for $r, s \geq 0$, combined with the reflecting boundary condition $\frac{1}{2}\frac{\partial P(r,\tau'|s,\tau)}{\partial r} + \delta P(r, \tau'|s, \tau) = 0$ for $r = 0$ and all $T > t$. This prevents a probability flow from $r > 0$ to $r < 0$. The initial condition is is $P(r, \tau'|s, \tau) \to \delta(r - s)$ as $\tau' \to \tau$. The Fokker-Planck equation is that of a diffusing particle under a constant gravitational force, where $r = 0$ is the surface of the earth acting as a reflecting boundary. The solution is found to be

$$\begin{aligned} P(r, \tau'|s, \tau) = & \frac{1}{\sqrt{2\pi\Delta\tau}} \exp\left(-\frac{(r - s + \delta\Delta\tau))^2}{2\Delta\tau)}\right) + \\ & \frac{e^{2s\delta}}{\sqrt{2\pi\Delta\tau)}} \exp\left(-\frac{(r + s + \delta\Delta\tau))^2}{2\Delta\tau)}\right) + \\ & \delta e^{-2r\delta}\left(1 - erf\left(\frac{r + s - \delta\Delta\tau)}{\sqrt{2\Delta\tau)}}\right)\right) \end{aligned} \tag{19}$$

with $\Delta\tau = \tau' - \tau$. This solution can be used to compute $\phi(s, t)$ for $s \geq 0$ via (15), ie.

$$\phi(s, \tau) = \int_0^\infty dr L(r) P(r, 1|s, \tau)$$

and ϕ is extended to negative s by setting $\phi(-s, t) = \phi(s, t)$. As an example we take the problem of a shepherd who tries to keep the sheep in an interval of size $2a$ corresponding to a loss $L_a \doteq I_{s>a}$, where I is the indicator function.

The following table contains the explicit results for ϕ for three loss functions. The variable $\theta \doteq 1 - \tau$.

Loss	$\phi(\mathbf{s}, \tau)$	sign$[w^*(s,\tau)]$
L_{BBM}	$\frac{1}{2}\left(1 - erf\left(\frac{s+\delta\theta}{\sqrt{2\theta}}\right)\right)$	1
L_e	$e^{c(s-\delta\theta)+\frac{1}{2}c^2 2\theta}$	-1
L_a	$\frac{1}{2}(1+e^{-2a\delta}) - \frac{1}{2}erf\left(\frac{a-s+\delta\theta}{\sqrt{2\theta}}\right)$ $-\frac{1}{2}e^{-2a\delta}erf\left(\frac{a+s-\delta\theta}{\sqrt{2\theta}}\right)$	-sign(s)

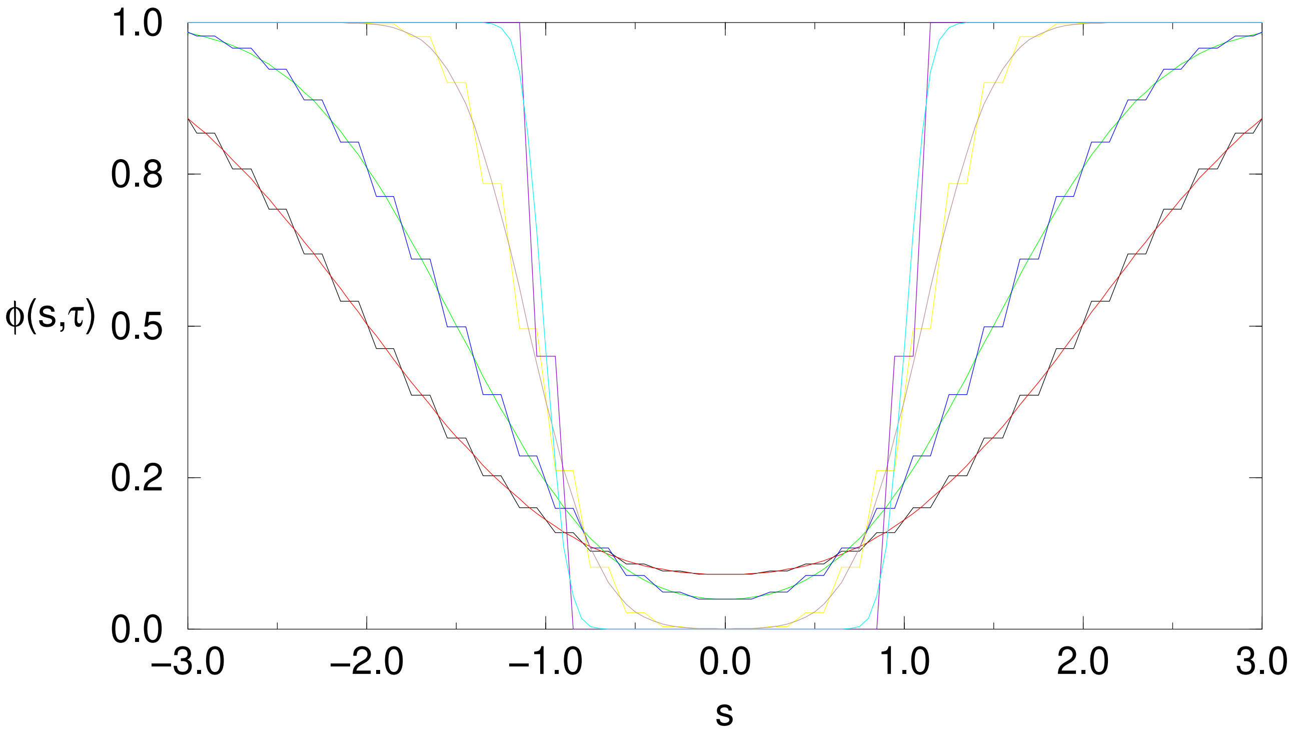

Figure 1: The potential $\phi(s,t)$ for the loss function $L_a = I_{s>a}$ as a function of $\mathbf{s}$ for $\delta = a = 1$ and (from left to middle) $\tau = 0, 0.5, 0.9, 0.99$. The step function is a result of a numerical iteration of Equation (5) with step size $\epsilon = 0.1$

The potential ϕ for the loss L_a is shown as the smooth curves in Fig. 1 for dif-

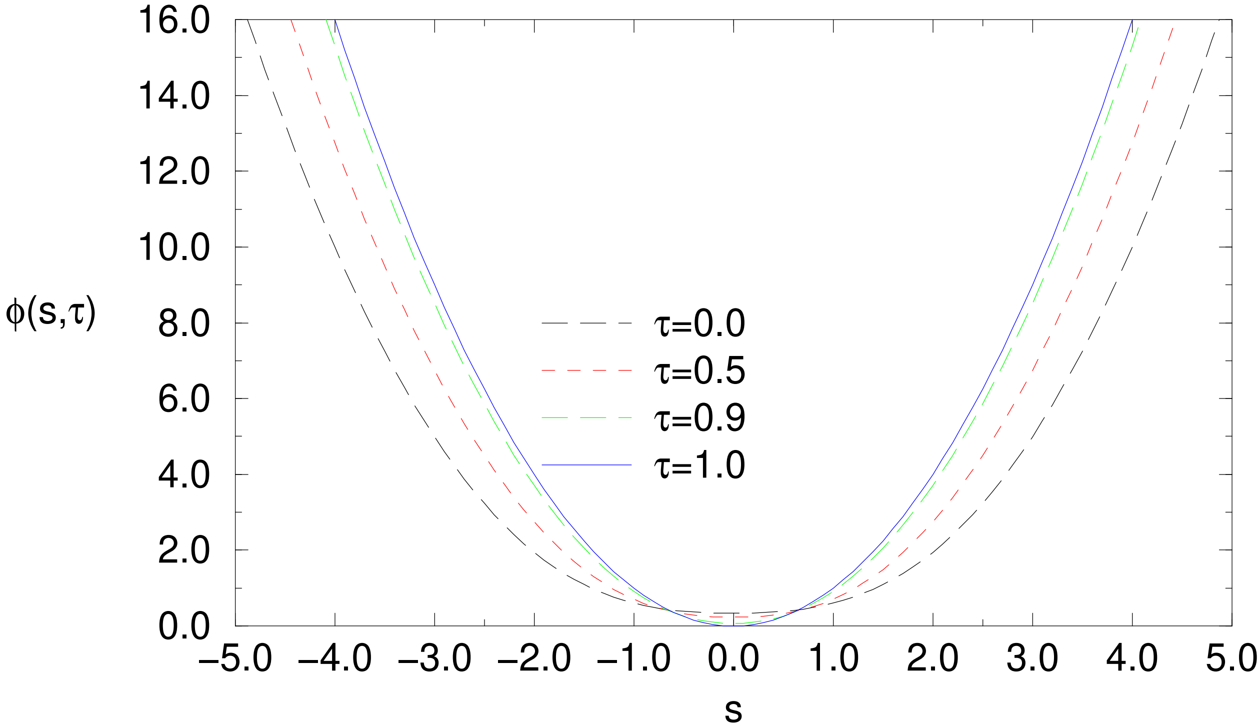

Figure 2: The potential $\phi(s,t)$ for the square loss $L(s) = s^2$.

ferent times. The step functions represent the corresponding solutions for Equation (5) with step size $\epsilon = 0.1$. We also computed ϕ for the two loss functions $L(s) = s^2$ and $L(y) = \min(s^2, 1)$ (see Figures 2 and 3) which maybe of interest in a regression framework. In these cases Equation (20) can still be expressed analytically in terms of error functions but the resulting expressions are long and complex. Instead, we have chosen to calculate the potential function by numerical integration.

9 Numerical solutions for $d = 2$

In higher dimensions as well as for loss functions which do not have the nice symmetries of the examples in the previous section we will have to resort to numerical solutions of the partial differential Equation (12).

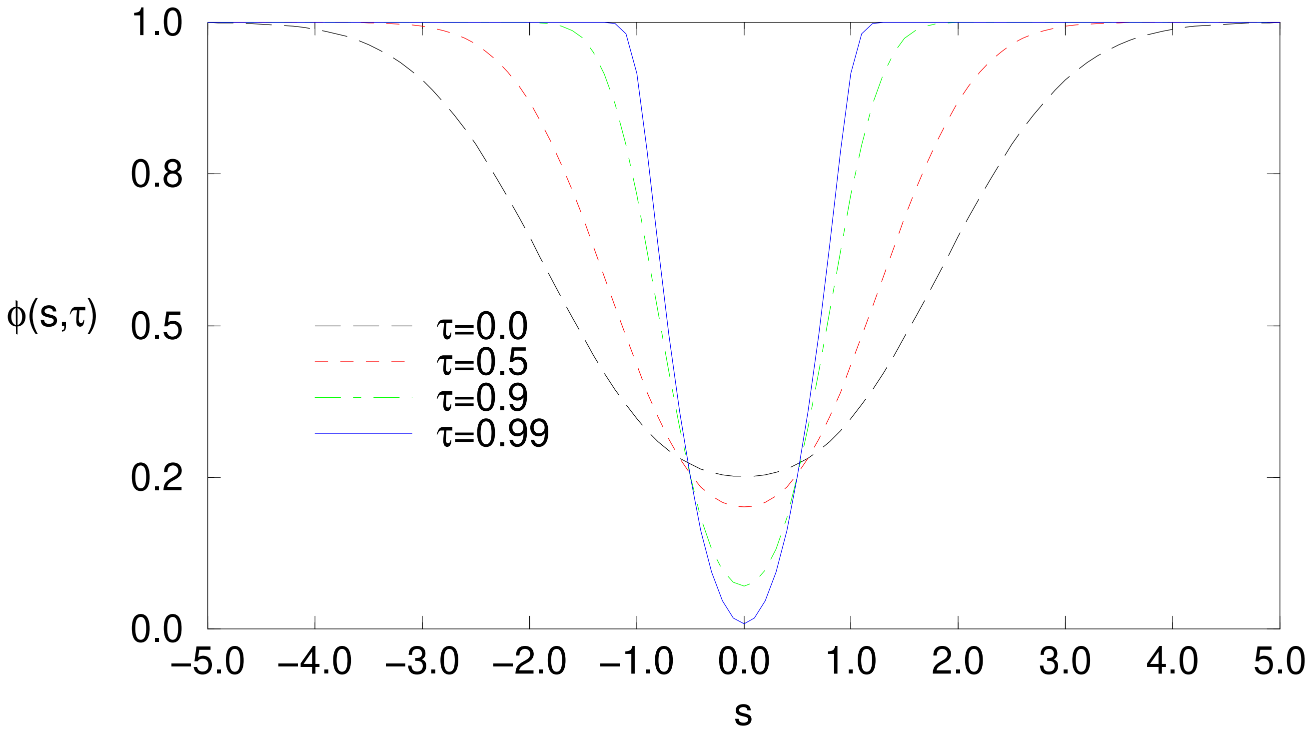

Figure 3: The potential $\phi(s,t)$ for the loss $L(s) = \min(s^2, 1)$.

To discuss some of the inherrent problems with such an approach we consider the regular lattice (4) in R^2. For this case, we have $D_{12} = D_{21} = 0$ and $D_{11} = D_{22} = \frac{1}{2}$. Numerical packages will usually solve PDEs *in forward time* starting from a given initial condition. Hence, we define the function $\psi(\mathrm{s}, t) = \phi(\mathrm{s}, 1-t)$ with the initial condition $\psi(\mathrm{s}, 0) = L(\mathrm{s})$. Writing the components of position vectors as $\mathrm{s} = (x, y)$ and choosing the L_2 norm, eq. (12) becomes

$$\frac{\partial\psi(\mathrm{s},t)}{\partial t} = \frac{1}{4}\frac{\partial^2\psi(\mathrm{s},t)}{\partial x^2} + \frac{1}{4}\frac{\partial^2\psi(\mathrm{s},t)}{\partial y^2} - \delta\sqrt{\left(\frac{\partial\psi(\mathrm{s},t)}{\partial x}\right)^2 + \left(\frac{\partial\psi(\mathrm{s},t)}{\partial y}\right)^2} . \quad (20)$$

To be specific, we will specialize to a loss function $L(\mathrm{s})$ which equals 1 if the point s is closest to $\mathbf{z}_1$ and 0 if it is closer to $\mathbf{z}_2$ or $\mathbf{z}_3$ than to $\mathbf{z}_1$. The region with $L(\mathrm{s}) = 1$ is a wedge of 60 degress width, centered at the origin $\mathrm{s} = 0$ and symmetric to the axis $x = 0$. This problem corresponds to boosting for classification problems with three possible labels as described by Schapire in [13]. The initial condition for the PDE reads

$$\psi(\mathrm{s}, 0) = I_{y>|x|/\sqrt{3}}. \quad (21)$$

Before attempting a numerical solution, we have to deal with two questions: How to choose appropriate *boundary conditions* and how to deal with the nonsmooth *initial conditions* given by (21).

Numerical solutions of the PDE must obviously be constrained to some finite region $D \subset R^2$. The choice of this region has to be supplemented by a sensible *a priori* specification of how the solution should behave at the artificial boundaries ∂D of the region D. Simply *fixing* $\psi(\mathrm{s}, t)$ for $\mathrm{s} \in \partial D$ for all times t to its initial values (corresponding to (21)) seems to create a rather crude approximation to the real problem defined on the entire R^2. We have rather chosen a boundary condition which models a situation where the diffusing “sheep” are not allowed to leave or enter the region D, i.e. the local *probability flow* at the boundary is required to be zero. Gardiner [10] shows that these *reflecting boundary* conditions as specified for the *backward equation* (12) are mathematically expressed as

$$\mathbf{n}(\mathrm{s}) \cdot \nabla\phi(\mathrm{s}, t) = 0 \qquad \text{for} \qquad \mathrm{s} \in \partial D\,. \quad (22)$$

$\mathbf{n}(\mathrm{s})$ is a vector perpendicular to the boundary at s [4].

The second problem that prevents us from a straightforward numerical solution of (20) is the discontinuity of the initial condition (21) at the line $y = |x|/\sqrt{3}$.

[4] (22) belongs to the class of “natural boundary conditions”.

This creates infinite spatial derivatives at time $t = 0$. Since it would require infinite precision in discretizing the PDE at this line, any *finite* spatial discretization may create uncontrollable artifacts and errors in the numerical solution. Hence, we have chosen to replace the original problem (21) by the smoothed version

$$\psi(\mathbf{s}, 0) = I^{(\lambda)}_{y>|x|/\sqrt{3}}, \tag{23}$$

which is based on a smoothed step function defined by $I^{(\lambda)}_x = \frac{1}{2}(1 + erf(x/\lambda))$. The parameter λ measures the typical lengthscale over which the step function smoothly varies from 0 to 1. The original I_x is recovered in the limit $\lambda \to 0$. This approach seems reasonable by the fact that the continuum limit of the PDE is understood as an approximation for a game with *finite* step size. Hence, a sensible choice is to take λ small compared to the step size of the original sheep game. We have solved the PDE (20,23) for $\delta = 1$ on the square $D = [-5, 5]^2$ using the *FLEXPDE* package[5]. We used a basic and straightforward definition of our problem as described in the code given in Appendix A. In order to obtain a reasonable approximation within the limits of the demo software we use the rather large smoothing parameter $\lambda = 1$ for the initial condition (23). The results for the potential $\phi(\mathbf{s}, t)$ for times $t = 1$ (the initial loss) and $t = 0$ are shown in Figure 4. The two straight lines in the $x - y$ plane show the function $y = |x|/\sqrt{3}$, where the unsmoothed loss function (21) jumps from $L = 0$ to $L = 1$. Since by symmetry, the potential comes out symmetric with respect to the axis $x = 0$ (i.e. $\phi(-x, y, t) = \phi(x, y, t)$), we have displayed $\phi(x, y, t = 1)$ for $x > 0$ and $\phi(x, y, t = 0)$ for $x < 0$ only. As expected from the relation (15), as the time t decreases, the potential becomes further smoothed. This effect is also visible in the contour plots (Figs. 5,6) where we have also shown the weight vectors (perpendicular to the lines of constant potential) as arrows. At $t = 0$, the potential is less steep than for $t = 1$.

A FLEXPDE code

```
TITLE ' brownboost'
VARIABLES
```

[5] *FLEXPDE* is a commercial PDE solver distributed by PDE-solutions Inc. It is based on a *finite element* approach which automatically creates a problem specific mesh over the spatial domain. Functions ψ are polynomially interpolated over each mesh cell. We used a demo version of FLEXPDE which can be downloaded from `http://www.pdesolutions.com`

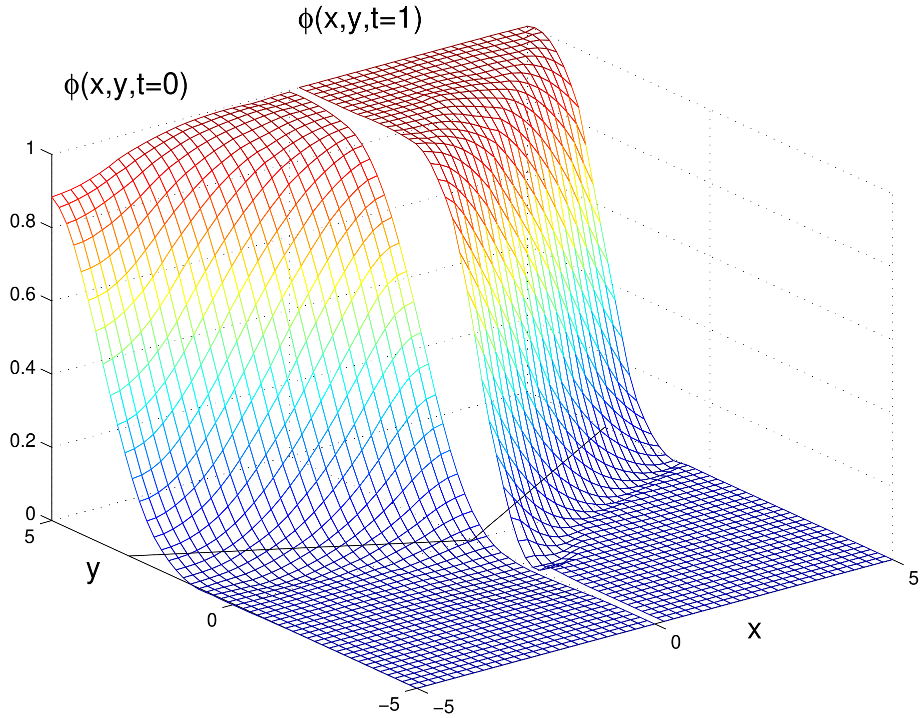

Figure 4: $\phi(\mathbf{s}, t)$ for $t = 1$ (only $x > 0$ shown) and $t = 0$ (only $x < 0$ shown) for the two dimensional problem. The missing parts for both functions follow by symmetry.

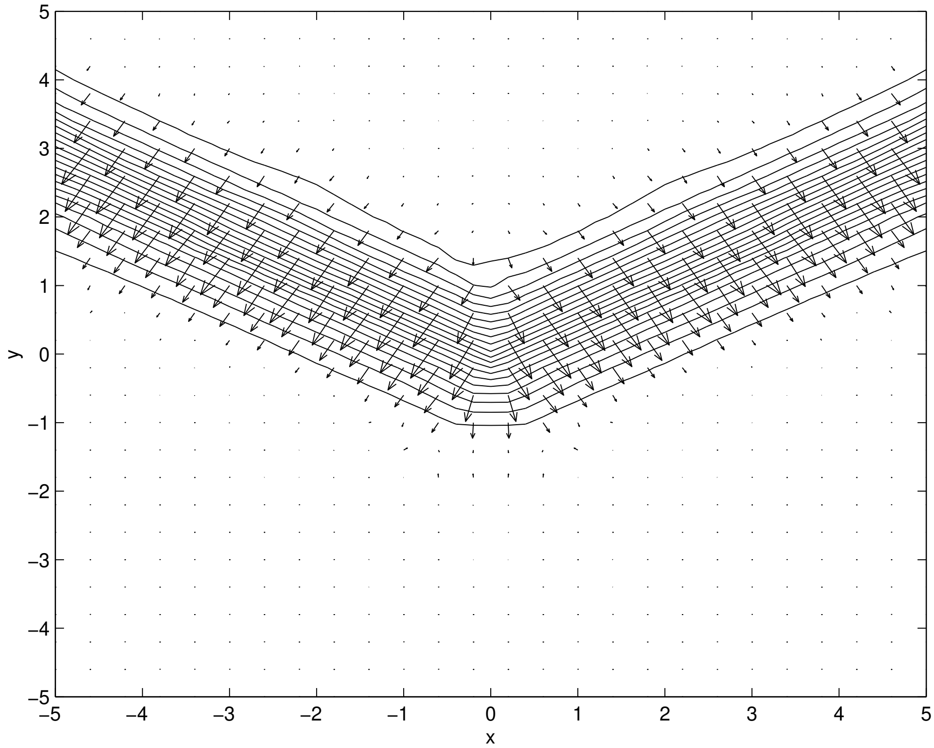

Figure 5: Contour plot of the potential (initial loss) $\phi(\mathbf{s}, t = 1)$ for the two dimensional problem together with the weightvectors $\mathbf{w}^*(\mathbf{s}, t = 1) = -\nabla(\mathbf{s}, t = 1)$.

```
Psi(range=0,1)
DEFINITIONS
lam=1.0
L=5
INITIAL VALUES
Psi=0.5*ERF((y-ABS(x)/sqrt(3))/lam)+0.5
EQUATIONS
dt(psi)= 0.25*div(grad(Psi))- Magnitude(grad(psi))
BOUNDARIES
region 1 'box'
start(-L,-L)
NATURAL(Psi)=0 line to (L,-L)
```

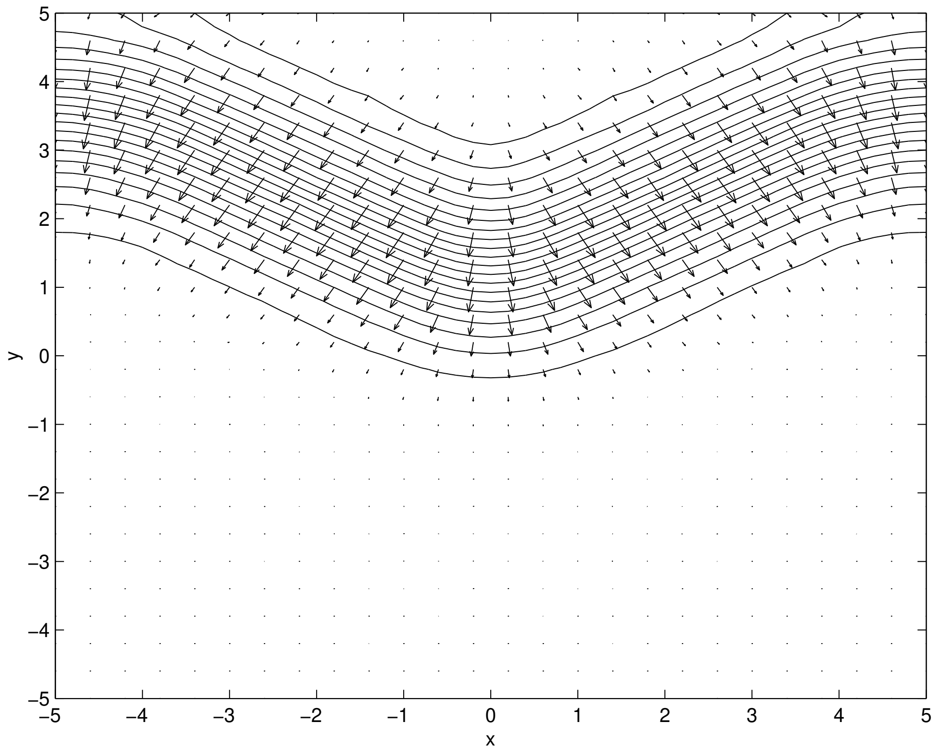

Figure 6: Contour plot of the potential $\phi(\mathbf{s}, t = 0)$ for the two dimensional problem together with the weightvectors $\mathbf{w}^*(\mathbf{s}, t = 0) = -\nabla(\mathbf{s}, t = 0)$.

```
NATURAL(Psi)=0 line to (L,L)
NATURAL(Psi)=0 line to (-L,L)
NATURAL(Psi)=0 line to finish
TIME
0 to 1 by 0.01
PLOTS
for t=0.0 by 0.5 to 1.0
CONTOUR(Psi) Gray
TABLE(Psi) Points = 51
END
```

References

[1] Javed Aslam. *Noise Tolerant Algorithms for Learning and Searching*. PhD thesis, Massachusetts Institute of Technology, 1995. MIT technical report MIT/LCS/TR-657.

[2] Javed A. Aslam and Aditi Dhagat. Searching in the presence of linearly bounded errors. In *Proceedings of the Twenty Third Annual ACM Symposium on Theory of Computing*, May 1991.

[3] Javed A. Aslam and Aditi Dhagat. On-line algorithms for 2-coloring hypergraphs via chip games. *Theoret. Comput. Sci.*, 112(2):355–369, 1993.

[4] Leo Breiman. *Probability*. SIAM, classics edition, 1992. Original edition first published in 1968.

[5] Nicolò Cesa-Bianchi, Yoav Freund, David P. Helmbold, and Manfred K. Warmuth. On-line prediction and conversion strategies. *Machine Learning*, 25:71–110, 1996.

[6] Aditi Dhagat, Péter Gács, and Peter Winkler. On playing "twenty questions" with a liar. In *Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms (Orlando, FL, 1992)*, pages 16–22, New York, 1992. ACM.

[7] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.

[8] Yoav Freund. An adaptive version of the boost by majority algorithm. *Machine Learning*, 43(3):293–318, June 2001.

[9] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[10] C.W. Gardiner. *Handbook of Stochastic Methods*. Springer Verlag, 2nd edition, 1985.

[11] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Functional gradient techniques for combining hypotheses. In Alexander J. Smola, Peter J. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*. MIT Press, 1999.

[12] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*, 2000.

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

[14] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, December 1999.

[15] Joel Spencer. *Ten Lectures on the Probabilistic Method*. Society for Industrial and Applied Mathematics, Philadelphia, 1987.

[16] Joel Spencer. Ulam's searching game with a fixed number of lies. *Theoret. Comput. Sci.*, 95(2):307–321, 1992.

[17] V. G. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2):153–173, April 1998.