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

Abstract

We combine the results of [5] and [3] 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 [2], 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 [5] 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 [3], 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 [5] and [3] 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 [5].

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 their solution using potential functions. 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 4 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 5 we take this limit and derive the differential equations that govern the game in this limit. In Section 6 we give a physical interpretation of Equations and the game. We conclude with some explicit solutions in Section 7.

2 Background

2.1 The Drifting Game

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 $\|\cdot\|_g$ 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.

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

$$\sum_{i=1}^{m} \mathbf{w}_i^t \cdot \zeta_i^t \geq \delta \sum_{i=1}^{m} \left\| \mathbf{w}_i^t \right\|_g \tag{1}$$

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 Analysis by Potential

In [5] 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:

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

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

One can show that the average potential is non increasing

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

Hence, one gets the bound on the average loss

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

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 $\sum_{i=0}^{d} z_i = 0$. We call a set B that satisfies these conditions 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 $i \in 0, \ldots, d$, $z_i \cdot z_i = a$
2. For any $i, j \in 0, \ldots, d$ such that $i \ne j$, $z_i \cdot z_j = -b$

For example, a regular set in R^2 is

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

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

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

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

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 $\delta_0 \ge \delta \ge 0$ the solution to the recursive definition of ϕ_{t-1} satisfies*

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

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

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

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

$$\delta_0 = \frac{1}{d}\min_{\mathbf{w}\ne 0}\frac{\left\|\mathbf{w}\right\|_2}{\left\|\mathbf{w}\right\|_g}$$

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 [5] 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

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

For some small $\mu > 0$.

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

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

As m is large and the steps are chosen independently at random the actual value of $\sum_{i=1}^{m}\mathbf{w}_i^t\zeta_i^t$ is likely to be very close to its expected value and thus, with high probability $\sum_{i=1}^{m}\zeta_i^t\mathbf{w}_i^t > \delta\sum_{i=1}^{m}\left\|\mathbf{w}_j^t\right\|_g$. 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

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

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 which implies that its minimum is achieved on an affine subspace (a translation of a linear subspace). We shall now show that this subspace consists of a single point.

To test whether a point $\mathbf{w}$ is a local minimum we consider the restriction of the function $F(\mathbf{w})$ on rays emanating from $\mathbf{w}$. Given a point $\mathbf{w}$ and a direction vector $\mathbf{v}$ such that $\|\mathbf{v}\|_g = 1$, 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 $g_{\mathbf{w},\mathbf{v}}(x) = x\max_i \mathbf{z}_i\mathbf{v}$. Thus $g_{\mathbf{w},\mathbf{v}}$ is constant if and only if $\max_i \mathbf{z}_i\mathbf{v} = 0$. Written in another way, this means that $\mathbf{z}_i\mathbf{v} \le 0$ for all $i = 0,\ldots,d$. Consider the two possibilities. If $\mathbf{z}_i\mathbf{v} = 0$ 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 $\mathbf{z}_i < 0$ then $\mathbf{v}\sum_i \mathbf{z}_i < 0$ 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 on 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

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

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

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

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 $D_{\mathbf{w},\mathbf{v}}(F) \ge 0$ 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 $\delta\|\mathbf{w}\|_g$.

We start with F_0. As shown earlier, the ray functions $\mathbf{g}_{\mathbf{w}^*,\mathbf{v}}$ for F_0 are equal to $g_{\mathbf{w},\mathbf{v}}(x) = x\max_i \mathbf{z}_i\mathbf{w}^*$. This implies two facts:

- For $\mathbf{w} = \mathbf{w}^*$ then $D_{\mathbf{w}^*,\mathbf{v}}(F_0) = \max_i \mathbf{z}_i\mathbf{v} > 0$ 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 potential function ϕ_t.
- For $\mathbf{w} \neq \mathbf{w}^*$ 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 $\mathbf{v} = (\mathbf{w}^*-\mathbf{w})/\|\mathbf{w}^*-\mathbf{w}\|_g$ and thus $D_{\mathbf{w},\mathbf{v}}(F_0) = -D_{\mathbf{w}^*,-\mathbf{v}}(F_0) < a < 0$.

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

Combining these two observations we conclude 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. This setting follows directly from observing that in this case the vectors $\mathbf{v}$ that minimizes $\max_i \mathbf{z}_i\mathbf{v}$ are $-\mathbf{z}_i$ and $\mathbf{z}_i\mathbf{z}_j/\|\mathbf{z}_i\|_2 = 1/d$ for any $i \neq j$ in a regular set of vectors. ■

4 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 [3], 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.

5 Continuum limit

We will now show that the latter definition of a continuum limit also leads to a natural limit of the recursion (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 $\mathbf{s}$, we expand the right hand side in a Taylor series up to second order in ϵ.

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

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

We introduce the continuous time variable τ via $t\epsilon^2$, and expand the left hand side of (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

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

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 $\mathbf{s}$ respectively. The linear term in ϵ vanishes due to the extra condition on the vectors $\mathbf{z}_i$. For the regular set described in (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 (6) to first order in ϵ. This simply yields the gradient

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

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

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

6 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 [1] 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 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. [1] $\mathbf{A}$ is usually called the *drift* field and D the Diffusion matrix. Taking the limit of small step size in (7) we get

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

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

[1] Note that on average the displacement (or velocity) is proportional to the Force. This behavior which is unlike the well known "acceleration $propto$ force" describes motion in a a viscous medium, where motion is strongly damped.

evolution of this conditional distribution is described by the *forward* or *Fokker-Planck* equation:

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

One can show that the PDE (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 expected loss at time $\tau' = 1$ when a sheep is at time τ at the position s i.e..

$$\phi(\mathbf{s},t) = \int d\mathbf{r} L(\mathbf{r}) p(\mathbf{r},\tau'=1|\mathbf{s},\tau) \quad (15)$$

By using the so called *Backward equation* ([4]) which describes the evolution $p(\mathbf{r},\tau'|\mathbf{s},\tau)$ with respect to the initial condition s and τ one arrives at (12).

7 Explicit solutions for $d = 1$

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

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

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

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, (16) is *linear*. We will discuss 2 cases next:

Monotonic loss: Here we have $\text{sign}[\frac{\partial L(s,\tau)}{\partial s}] = const$ for all $s \in R$ leading to $\text{sign}[w^*(s,\tau)] = const$. Special examples are "Boost by Majority" loss $L_{BBM}(x) = 1$ for $x < 0$ and $L_{BBM}(x) = 0$ else. and the exponential loss $L_e(y) = e^{cy}$.

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 setting $A(s,\tau) = \text{sign}(w^*(s,\tau))$. We illustrate this for the case of *increasing* loss functions $A(s,\tau) = 1$ for $s \geq 0$.

$$\frac{\partial P(r,\tau'|s,\tau)}{\partial\tau'} = \frac{1}{2}\frac{\partial^2 P(r,\tau'|s,\tau)}{\partial r^2} + \delta\frac{\partial P(r,\tau'|s,\tau)}{\partial r} \quad (17)$$

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

$$P(r,\tau'|s,\tau) = \frac{1}{\sqrt{2\pi\Delta\tau}}\exp\left(-\frac{(r-s+\delta\Delta\tau))^2}{2\Delta\tau)}\right) + \frac{e^{2s\delta}}{\sqrt{2\pi\Delta\tau)}}\exp\left(-\frac{(r+s+\delta\Delta\tau))^2}{2\Delta\tau)}\right) + \delta e^{-2r\delta}\left(1 - erf\left(\frac{r+s-\delta\Delta\tau)}{\sqrt{2\Delta\tau)}}\right)\right) \quad (18)$$

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

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

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 $L_a \doteq I_{x>a}$, 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	$\phi(\mathbf{s},\tau)$	$\text{sign}[w^*(s,\tau)]$
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)

The potential ϕ for the loss L_a is shown as the smooth curves in Fig. 1 for different times. The step functions present the corresponding solutions for the discrete recursion (5) with a step size $\epsilon = 0.1$. We also computed ϕ for the two loss functions $L(y) = y^2$ and $L(y) = \min(y^2, 1)$ (see Figs. 2 and 3) which maybe of interest in a regression framework. Although in these cases (19) may still be evaluated in terms of error functions in a complicated way, we have rather evaluated (19) by numerical integration instead.

References

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

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

[3] Yoav Freund. An adaptive version of the boost by majority algorithm. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 1999.

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

[5] Robert E. Schapire. Drifting games. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 1999.

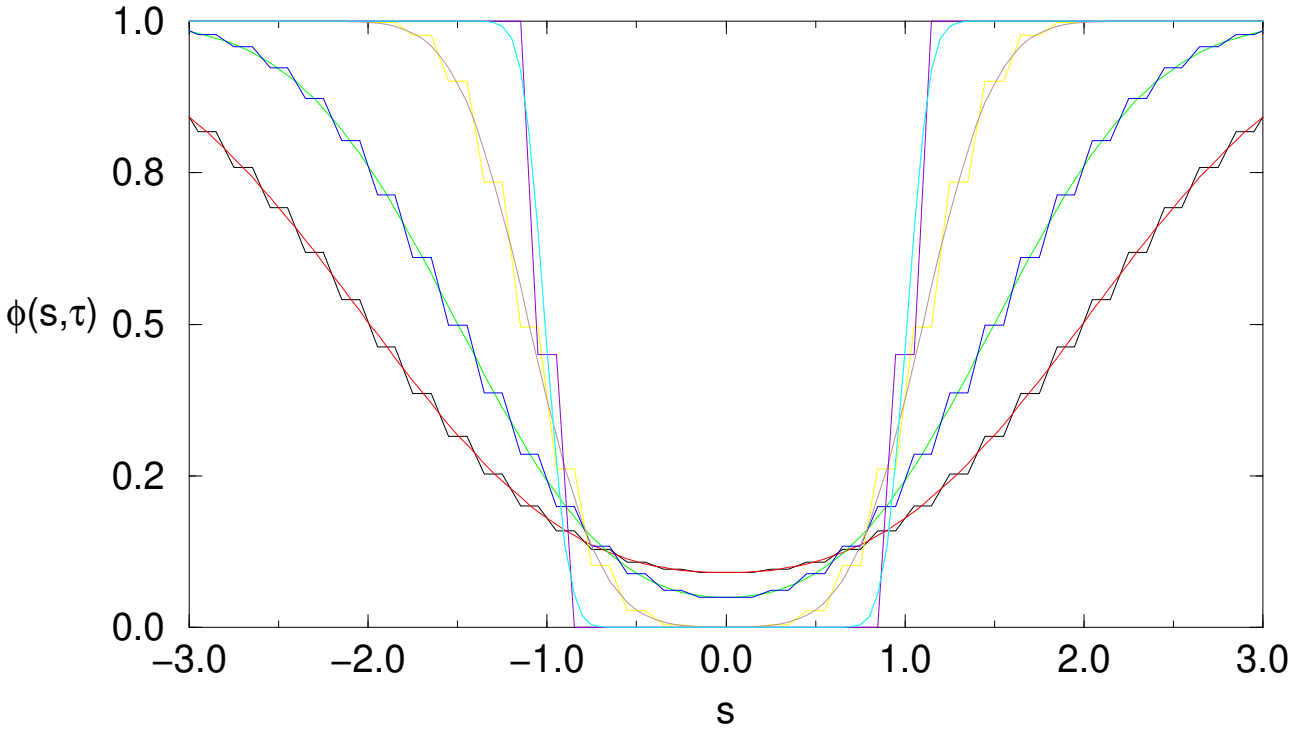

Figure 1: The potential $\phi(\mathrm{s}, t)$ for the loss function $L_a = I_{y>a}$ as a function of s for $\delta = a = 1$ and (from left to middle) $\tau = 0, 0.5, 0.9, 0.99$. The step function is a result of a numerical iteration of the discrete recursion (5) with step size $\epsilon = 0.1$

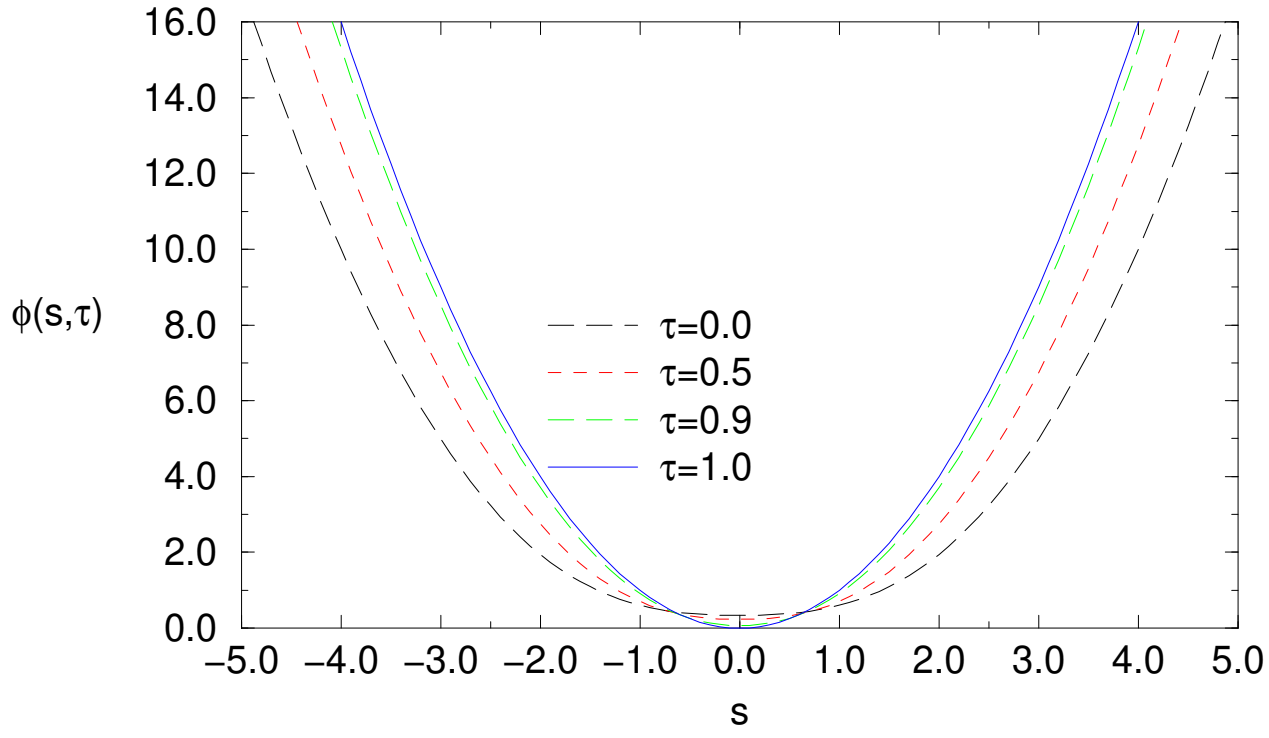

Figure 2: The potential $\phi(\mathrm{s}, t)$ for the square loss $L(y) = y^2$.

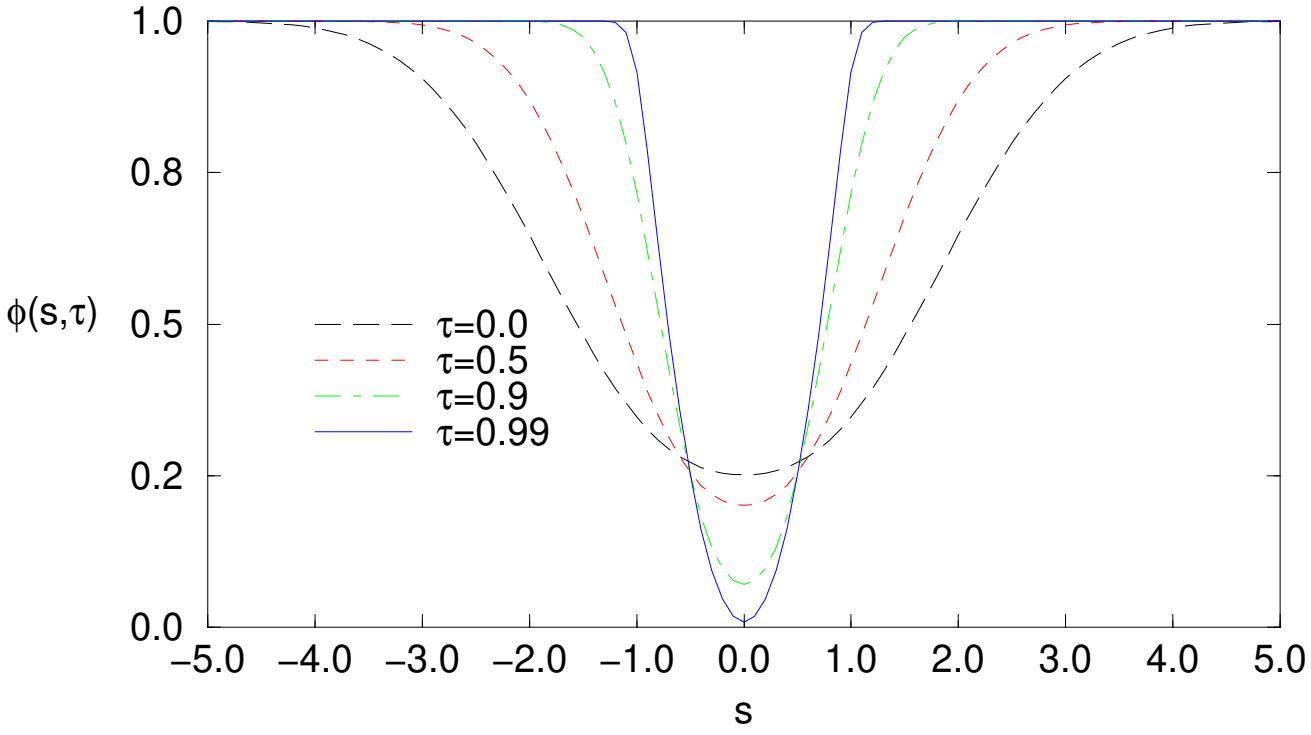

Figure 3: The potential $\phi(\mathrm{s}, t)$ for the loss $L(y) = \min(y^2, 1)$.