14 Boosting in Continuous Time

AdaBoost owes much of its practicality as a boosting algorithm to its adaptiveness, its ability to automatically adjust to weak hypotheses with varying accuracies, thus alleviating the need for knowledge prior to the start of the boosting process of a minimum edge γ, or even the total number of rounds T that will be run. The boost-by-majority algorithm, studied in chapter 13, is not adaptive. Still, it may have other advantages over AdaBoost: It is theoretically more efficient (in terms of the number of rounds to achieve some accuracy) and, perhaps more importantly, it may be better at handling outliers. In this companion chapter, we study a method for making BBM adaptive while, hopefully, retaining its other positive qualities.

BBM is actually nonadaptive in two senses. First, it requires knowledge of a value $\gamma > 0$ for which the γ-weak learning assumption holds so that all weak hypotheses have edge at least γ. And second, it is nonadaptive in its inability to fully exploit weak hypotheses which happen to have edges significantly better than γ. We will see that this latter form of nonadaptiveness can be overcome by adjusting the weights on the weak hypotheses and allowing the algorithm's "clock" to advance by more than one "tick" on each round. However, the resulting algorithm still requires knowledge of the minimum edge γ. To handle this difficulty, we imagine allowing γ to become very small, while simultaneously increasing the total number of rounds T. In the limit $\gamma \to 0$, the number of rounds becomes infinite. If the total time of the entire boosting process is nevertheless squeezed into a finite interval, then in the limit, boosting is conceptually proceeding in *continuous* time. The result is a continuous-time version of BBM called BrownBoost that, like AdaBoost, can adapt to varying edges among the weak hypotheses.

In section 13.3.2, we saw that NonAdaBoost, a nonadaptive version of AdaBoost, can be derived from BBM. Correspondingly, we will see in this chapter that AdaBoost, in its usual, adaptive form, is itself a special case of BrownBoost. Or, in other words, BrownBoost is a generalization of AdaBoost. As will be seen, this generalization explicitly incorporates an anticipated *in*ability to drive the training error to zero, as is to be expected with noisy data, or data containing outliers.

We end this chapter with some experiments comparing BrownBoost and AdaBoost on noisy data.

14.1 Adaptiveness in the Limit of Continuous Time

Our goal is to make BBM adaptive. As noted above, its non-adaptiveness takes two forms, namely, required prior knowledge of a minimum edge γ, together with an inability to fully exploit weak hypotheses with edges much better than γ. We begin with an informal overview of the main ideas for overcoming each of these.

14.1.1 Main Ideas

Suppose on some round t that a weak hypothesis h is received from the weak learner with weighted error substantially below $\frac{1}{2} - \gamma$, in other words, with edge much larger than the minimum requirement of γ. Even when this happens, BBM will treat h like any other weak hypothesis, essentially ignoring its relative strength. Thus, h will be used just once, and an entirely new weak hypothesis will be sought on the following round.

There is, however, a natural alternative. Under the conditions above, it may well happen that h's error continues to be smaller than $\frac{1}{2} - \gamma$ when measured with respect to the *new* distribution D_{t+1}. In this case, h can be used a second time on round $t + 1$, just as if it had been received fresh from the weak learner. This may happen yet again on the following round, so that h can be used a third time. And continuing in this way, the same weak hypothesis h may be used many times until at last its error exceeds $\frac{1}{2} - \gamma$. At this point, a new weak hypothesis must be obtained from the weak learner, and the process begins again. In the end, weak hypotheses with edges significantly exceeding γ will be included many times in the majority-vote classifier formed by BBM so that this final hypothesis will actually be a *weighted* majority over the weak hypotheses, with the most weight assigned to weak hypotheses with the lowest weighted error, just like AdaBoost. This already can be seen to be a form of adaptiveness.

This idea can be understood and generalized by considering the potential function $\Phi_t(s)$ at the heart of BBM, as studied in detail in section 13.1. Recall that the essence of our analysis of BBM's training error was theorem 13.3, a proof that the total (or average) potential of the m chips (or training examples) never increases from round to round. Since the final average potential is exactly the training error, this implied an immediate bound on the training error in terms of the initial potential $\Phi_0(0)$, as seen in corollary 13.4. Thus, a given desired training error of $\epsilon > 0$ can be attained simply by choosing the number of rounds T large enough that $\Phi_0(0) \leq \epsilon$.

In fact, this proof technique permits great freedom in how we use a given weak hypothesis, provided that the total potential is not allowed to increase. Given a weak hypothesis h, BBM simply increments the "clock" t:

$t \leftarrow t + 1,$

and advances the position s_i of each chip i by $z_i \doteq y_i h(x_i)$:

$$s_i \leftarrow s_i + z_i .$$

But there are other possibilities in how these might be updated. As seen above, we can use the same weak hypothesis h many times, say for k consecutive rounds. This is equivalent to advancing the clock t by k:

$$t \leftarrow t + k,$$

and advancing the chips by k times their usual increment of z_i:

$$s_i \leftarrow s_i + k z_i .$$

Under the assumption that h has weighted error at most $\frac{1}{2} - \gamma$ on each of the k time steps, theorem 13.3 implies that the total potential at the end of these k steps will be no larger than at the beginning. The point, however, is that this is the *only* property we care about for the analysis.

This observation opens the door to immediate generalization. As a start, we can decouple the amount by which the clock and the chips are advanced so that the clock is advanced, say, by some positive integer ξ:

$$t \leftarrow t + \xi,$$

and the chips by some integer increment α:

$$s_i \leftarrow s_i + \alpha z_i ,$$

where we no longer require $\xi = \alpha$. In fact, we can allow any choice of ξ and α, so long as the total potential does not increase. We do not here specify particular choices, but intuitively, we may wish to choose ξ large to speed the entire process which must end when the clock t reaches T.

Suppose, on the r-th round of this process, that a weak hypothesis h_r is received, and the clock and chips are advanced by ξ_r and α_r as above. Then the final hypothesis will be the weighted majority vote

$$H(x) \doteq \operatorname{sign}\left(\sum_{r=1}^{R} \alpha_r h_r(x)\right),$$

where R is the total number of rounds until the clock reaches T (and where we now carefully distinguish between "rounds" r and "time steps" t). Under this definition, an example (x_i, y_i) is misclassified by H if and only if the corresponding chip has been moved by the process described above to a final position z_i that is not positive. Thus, by exactly the same proof as in corollary 13.4, the training error of H can be shown to be at most the initial potential $\Phi_0(0)$, provided we respect the requirement that the total potential must never increase.

In this way, BBM can be modified to exploit weak hypotheses with varying edges, provided they are all at least γ. This latter condition, of course, remains a serious obstacle. A natural idea to get around it is simply to choose γ so small that it is almost sure to fall below the edges of all the weak hypotheses. In the limit $\gamma \to 0$, this is certain to be the case. Of course, according to our analysis of BBM, to achieve the same accuracy in the final classifier with smaller values of γ requires a correspondingly larger number of time steps T. Thus, as $\gamma \to 0$, T becomes infinite. If we rescale our notion of time, holding it fixed within some finite interval, this will mean in the limit that time is advancing *continuously* rather than in discrete steps.

So, to summarize, to remove the assumption of γ-weak learnability, we consider the continuous-time limit of BBM obtained by letting $\gamma \to 0$, combined with the technique given above for handling weak hypotheses with varying edges. To implement these ideas, we will first need to derive the continuous-time limits of both the potential function $\Phi_t(s)$ and the weighting function $w_t(s)$. Furthermore, we will need a technique for computing how much the clock and the chips should be advanced so as to maximize the progress that can be wrung from each weak hypothesis, subject to the condition that the average potential should never increase.

We turn now to a detailed treatment.

14.1.2 The Continuous-Time Limit

Our initial goal is to understand how the various elements of BBM behave in the limit as $\gamma \to 0$, and as the number of time steps T simultaneously grows to infinity. In the usual setting for BBM, "time" is indexed by integers $t = 0, 1, \ldots, T$, and similarly, "space"—that is, the positions of the chips—is indexed by integers $s \in \mathbb{Z}$. Thus, both time and space are discrete, and the potential function $\Phi_t(s)$ and the weighting function $w_t(s)$ are defined in terms of these discrete quantities.

Now, as we let T get large, it will be natural to focus not on the *actual* number of time steps t of BBM that have elapsed, but rather on the *fraction* of the T time steps that have passed, which we denote by

$$\tau = \frac{t}{T}. \tag{14.1}$$

In other words, it makes sense to rescale our notion of time so that boosting begins at time $\tau = 0$ and ends at time $\tau = 1$. Each discrete time step of BBM then takes, after rescaling, time $1/T$. As T increases, this tiny increment approaches zero, at which point boosting is conceptually proceeding in continuous time.

We will see shortly that our notion of space also will become continuous so that at each (continuous) moment in time $\tau \in [0, 1]$, each chip will occupy a continuously valued position $\psi \in \mathbb{R}$. The potential function $\Phi_t(s)$, which measures the potential at each chip position at each moment in time, must thus be correspondingly replaced by a function $\Phi(\psi, \tau)$ that

Table 14.1
Some key quantities used in the (discrete-time) derivation of BBM, and their continuous-time analogues

	BBM	Continuous Time
time	t	τ
margin/chip position	s	ψ
potential function	$\Phi_t(s)$	$\Phi(\psi,\ \tau)$
weighting function	$w_t(s)$	$w(\psi,\ \tau)$

is defined in terms of these continuous variables, and that is itself the limit, after appropriate rescaling, of $\Phi_t(s)$. Similarly for the weighting function. (For notational reference, table 14.1 summarizes some key quantities for BBM and their continuous-time analogues.)

We have already noted that we require a limit in which $\gamma \to 0$ as $T \to \infty$. In fact, for this limit to be meaningful, it will be necessary that the values of T and γ be coupled appropriately to one another. Specifically, we have seen that the training error of BBM is at most the tail of the binomial distribution given in equation (13.37), which is approximately $e^{-2\gamma^2 T}$ by Hoeffding's inequality. Thus, for this bound to have a meaningful finite limit, we need the product $\gamma^2 T$ to be held fixed. To do this, we let $T \to \infty$, and set

$$\gamma = \frac{1}{2}\sqrt{\frac{\beta}{T}} \tag{14.2}$$

where β is a constant whose value we discuss later. (The factor of $\frac{1}{2}$ has no real impact since β is an arbitrary constant.) Clearly, this choice of γ converges to zero, while $\gamma^2 T$ is held to the fixed constant $\beta/4$.

The next step is to determine the limits of the weighting and potential functions which are at the foundation of BBM. The weighting functions plotted in figure 13.4 (p. 452) strongly resemble normal distributions. This is because they are binomial distributions which are well known to converge to normal distributions. To compute their limits precisely, recall from equation (13.38) that the potential $\Phi_t(s)$ turns out to be exactly equal to the probability that a particular random walk on the set of integers $\mathbb{Z}$, beginning at s, will end at a nonpositive value. Specifically, equation (13.38) can be rewritten as

$$\Phi_t(s) = \mathbf{Pr}\left[s + Y_{\overline{T}} \le 0\right] = \mathbf{Pr}\left[Y_{\overline{T}} \le -s\right] \tag{14.3}$$

where

$$\overline{T} \doteq T - t = T(1-\tau), \tag{14.4}$$

and where

$$Y_{\overline{T}} = \sum_{j=1}^{\overline{T}} X_j$$

is a sum of independent random variables X_j, each of which is $+1$ with probability $\frac{1}{2}+\gamma$, and -1 otherwise. The central limit theorem tells us that such a sum of independent random variables, if appropriately scaled and translated, will converge in distribution to a normal distribution as $\overline{T} \to \infty$. (See appendix A.9 for further background.) In this case, the mean of the sum $Y_{\overline{T}}$ is $2\gamma\overline{T}$, and its variance is $(1-4\gamma^2)\overline{T}$. Thus, subtracting the mean and dividing by the standard deviation gives the standardized sum

$$\frac{Y_{\overline{T}} - 2\gamma\overline{T}}{\sqrt{(1-4\gamma^2)\overline{T}}}, \tag{14.5}$$

which, by the central limit theorem, converges as $\overline{T} \to \infty$ to a standard normal distribution with mean 0 and unit variance.

As γ gets very small, its appearance in the denominator of equation (14.5) becomes negligible. Thus, for $\overline{T}$ large, equation (14.5) can be approximated by

$$\tilde{Y}_{\overline{T},\gamma} \doteq \frac{Y_{\overline{T}} - 2\gamma\overline{T}}{\sqrt{\overline{T}}} = \frac{Y_{\overline{T}}}{\sqrt{T(1-\tau)}} - \sqrt{\beta(1-\tau)}$$

by equations (14.2) and (14.4). Since this random variable is asymptotically the same as equation (14.5) (each differing from the other by a factor that converges to 1), its distribution also converges to standard normal.

The event $Y_{\overline{T}} \leq -s$ appearing in equation (14.3) holds if and only if

$$\tilde{Y}_{\overline{T},\gamma} \leq -\frac{s}{\sqrt{T(1-\tau)}} - \sqrt{\beta(1-\tau)}. \tag{14.6}$$

We would like the quantity on the right not to depend explicitly on T so that its limit will be meaningful. This can be achieved in the way that the discrete positions s of chips are replaced by continuous positions ψ, an operation we alluded to earlier but did not specify. Now we can be precise and define the linear mapping from discrete to continuous positions that we will use:

$$\psi = s\sqrt{\frac{\beta}{T}}. \tag{14.7}$$

Here, a scaling factor proportional to $1/\sqrt{T}$ has been chosen for the purpose of "absorbing" the appearance of this same factor on the right-hand side of equation (14.6); in particular, this definition causes the quantity $s/\sqrt{T}$ which appears in that expression now to be replaced simply by $\psi/\sqrt{\beta}$. (The constant $\sqrt{\beta}$ in equation (14.7) is arbitrary, and was chosen for mathematical convenience in what follows.) Thus, with ψ defined as above, the right-hand side of equation (14.6) can now be written as

$$-\frac{\psi}{\sqrt{\beta(1-\tau)}} - \sqrt{\beta(1-\tau)} = -\frac{\psi + \beta(1-\tau)}{\sqrt{\beta(1-\tau)}}. \tag{14.8}$$

Let Y^* be a standard normal random variable (with zero mean and unit variance). By the argument given above, $\Phi_t(s)$ is equal to the probability that $\tilde{Y}_{\overline{T},\gamma}$ is at most equation (14.8), which converges to

$$\Phi(\psi,\ \tau) \doteq \mathbf{Pr}\left[Y^* \leq -\frac{\psi + \beta(1-\tau)}{\sqrt{\beta(1-\tau)}} \right].$$

To summarize, we have shown that for any ψ and τ, if s and t are chosen to satisfy the scaling given in equations (14.1) and (14.7) (or to nearly satisfy these equations, given that they must be integers), and if γ is chosen as in equation (14.2), then as $T \to \infty$, the potential function $\Phi_t(s)$, which depends implicitly on T and γ, converges to $\Phi(\psi,\ \tau)$. In other words, $\Phi(\psi,\ \tau)$, under appropriate rescaling of the relevant variables, is the limit of BBM's potential function.

By the definition of the normal distribution, $\Phi(\psi,\ \tau)$ can be defined equivalently in a form that does not reference the random variable Y^*, namely,

$$\Phi(\psi,\ \tau) \doteq \frac{1}{2}\mathrm{erfc}\left(\frac{\psi + \beta(1-\tau)}{\sqrt{2\beta(1-\tau)}} \right) \tag{14.9}$$

where $\mathrm{erfc}(u)$ is the *complementary error function*

$$\mathrm{erfc}(u) \doteq \frac{2}{\sqrt{\pi}} \int_u^\infty e^{-x^2} dx, \tag{14.10}$$

which is plotted in figure 14.1. Thus, we have arrived at a closed-form expression for the limit of the potential function.

14.1.3 An Alternative Derivation

Although this derivation is complete, we here present a rather different method for deriving the potential function $\Phi(\psi,\ \tau)$, one based on setting up and solving a partial differential equation. We start from scratch with equation (13.29), the recursive formulation of the potential function, which, by direct substitution, implies that

$$\begin{aligned} \Phi_t(s) - \Phi_{t-1}(s) &= \Phi_t(s) - \left[\left(\tfrac{1}{2} + \gamma\right) \Phi_t(s+1) + \left(\tfrac{1}{2} - \gamma\right) \Phi_t(s-1) \right] \\ &= -\tfrac{1}{2} \left(\Phi_t(s+1) - 2\Phi_t(s) + \Phi_t(s-1) \right) \\ &\quad - \gamma \left(\Phi_t(s+1) - \Phi_t(s-1) \right). \end{aligned} \tag{14.11}$$

Next, we rewrite this equation in the continuous domain in terms of $\Phi(\psi,\ \tau)$. For the moment, we identify $\Phi(\psi,\ \tau)$ with $\Phi_t(s)$, where $\tau = t/T$ and $\psi = s\sqrt{\beta/T}$ as before, so that $\Phi(\psi,\ \tau)$ depends implicitly on T, a dependence that will vanish when the limit $T \to \infty$ is taken. As noted earlier, every step of BBM causes τ to increase by $\Delta\tau \doteq 1/T$.

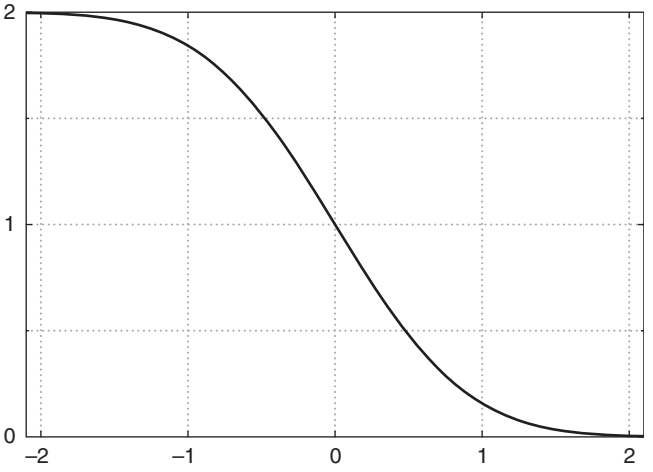

Figure 14.1
A plot of the function $\operatorname{erfc}(u)$ as given in equation (14.10).

In addition, when s is incremented or decremented by 1, ψ, by its definition in terms of s, is incremented or decremented by

$$\Delta\psi \doteq \sqrt{\frac{\beta}{T}} = \sqrt{\beta\Delta\tau} = 2\gamma$$

by equation (14.2). Plugging in these notational changes, equation (14.11) becomes

$$\begin{aligned}\Phi(\psi,\,\tau) - \Phi(\psi,\,\tau-\Delta\tau) = &-\tfrac{1}{2}\left[\Phi(\psi+\Delta\psi,\,\tau) - 2\Phi(\psi,\,\tau) + \Phi(\psi-\Delta\psi,\,\tau)\right] \\ &-\gamma\left[\Phi(\psi+\Delta\psi,\,\tau) - \Phi(\psi-\Delta\psi,\,\tau)\right].\end{aligned}$$

Dividing both sides by

$$-\beta\Delta\tau = -(\Delta\psi)^2 = -2\gamma\Delta\psi,$$

we arrive at the following difference equation:

$$\begin{aligned}-\frac{1}{\beta}\cdot\frac{\Phi(\psi,\,\tau)-\Phi(\psi,\,\tau-\Delta\tau)}{\Delta\tau} = &\frac{1}{2}\cdot\frac{\Phi(\psi+\Delta\psi,\,\tau)-2\Phi(\psi,\,\tau)+\Phi(\psi-\Delta\psi,\,\tau)}{(\Delta\psi)^2} \\ &+\frac{\Phi(\psi+\Delta\psi,\,\tau)-\Phi(\psi-\Delta\psi,\,\tau)}{2\Delta\psi}.\end{aligned} \tag{14.12}$$

Taking the limit as $T\to\infty$, so that $\Delta\tau\to 0$ and $\Delta\psi\to 0$, gives the following partial differential equation:

$$-\frac{1}{\beta}\cdot\frac{\partial\Phi(\psi,\,\tau)}{\partial\tau} = \frac{1}{2}\cdot\frac{\partial^2\Phi(\psi,\,\tau)}{\partial\psi^2} + \frac{\partial\Phi(\psi,\,\tau)}{\partial\psi}. \tag{14.13}$$

To derive this, we used the fact that for any differentiable function $f : \mathbb{R} \to \mathbb{R}$,

$$\frac{f(x+\Delta x) - f(x)}{\Delta x}$$

converges, in the limit $\Delta x \to 0$, to $f'(x)$, the derivative of f at x, by definition. We also used the fact that

$$\frac{f(x+\Delta x) - 2f(x) + f(x-\Delta x)}{(\Delta x)^2} = \frac{\frac{f(x+\Delta x)-f(x)}{\Delta x} - \frac{f(x)-f(x-\Delta x)}{\Delta x}}{\Delta x}$$

converges to $f''(x)$, the second derivative of f, as $\Delta x \to 0$.

Thus, in the limit, $\Phi(\psi, \tau)$ must satisfy equation (14.13). This equation turns out to be well known: It describes the time evolution of a so-called Brownian process, which is the continuous-time limit of a random walk.

Recall that at the end of a run of BBM, at time T, the potential function $\Phi_T(s)$ is defined to be an indicator function that counts training mistakes as in equation (13.25). Therefore, in the continuous-time limit, the potential function at the end of the boosting process, $\tau = 1$, should satisfy

$$\Phi(\psi, 1) = \mathbf{1}\{\psi \le 0\}. \tag{14.14}$$

This equation acts as a kind of "boundary condition." Solving the partial differential equation in equation (14.13) subject to equation (14.14) gives exactly equation (14.9), as can be verified by plugging the solution into the equation (see exercise 14.1). Thus, we have obtained the same limiting potential function as before.

As a technical point, we note that $\Phi(\psi, \tau)$ is continuous on its entire range, except at the point $\psi = 0, \tau = 1$. A discontinuity at this point is inevitable. And although equation (14.14) defines $\Phi(0, 1)$ to be 1, it could perhaps more reasonably be defined to be $\frac{1}{2}$. We discuss this annoying discontinuity further below, including how to stay away from it.

The weighting function $w_t(s)$ also gets replaced by a function $w(\psi, \tau)$ in terms of the new continuous variables. Since multiplying the weights by a positive constant has no effect, due to normalization, we divide the formula for $w_t(s)$ given in equation (13.33) by $\Delta\psi = \sqrt{\beta/T}$, so that $\sqrt{T/\beta} \cdot w_t(s)$ becomes

$$\sqrt{\frac{T}{\beta}} \cdot \frac{\Phi_t(s-1) - \Phi_t(s+1)}{2} = \frac{\Phi(\psi - \Delta\psi, \tau) - \Phi(\psi + \Delta\psi, \tau)}{2\Delta\psi}.$$

In the limit $\Delta\psi \to 0$, this gives the weighting function

$$w(\psi, \tau) = -\frac{\partial\Phi(\psi, \tau)}{\partial\psi} \propto \exp\left(-\frac{(\psi + \beta(1-\tau))^2}{2\beta(1-\tau)}\right), \tag{14.15}$$

where we write $f \propto g$ to mean that f is equal to g times a positive constant that does not depend on ψ.

Both the potential function $\Phi(\psi, \tau)$ and the weighting function $w(\psi, \tau)$ are plotted for sample values of τ in figure 14.2. Since these are the limits of the corresponding functions for BBM, it is not surprising that the weighting function in this figure is almost identical to the one shown in figure 13.4 for BBM with $T = 1000$ (other than being a lot smoother).

14.2 BrownBoost

Having computed the limit of the potential and weighting functions, we can return to our earlier ideas for the design of an adaptive boost-by-majority algorithm.

14.2.1 Algorithm

Given our usual dataset of m training examples, the state of the continuous-time algorithm can be described by the current time $\tau \in [0, 1]$, and by the position ψ_i of each chip/training example i. From the derivation above, we can compute weights $w(\psi_i, \tau)$ for each of these which, when normalized, define a distribution D. The weak learning algorithm can be used to obtain a weak hypothesis h whose error with respect to D is less than $\frac{1}{2}$, as usual. What, then, do we do with h?

According to the outline of ideas presented in section 14.1.1, the clock and chip positions should now be advanced by some amounts, subject to various conditions. Specifically, applying these earlier ideas to the newly derived continuous-time domain, the clock should be advanced by some amount $\xi > 0$ from τ to $\tau' = \tau + \xi$, and each chip i should be moved in the direction $y_i h(x_i)$ by some amount α to the new position

$$\psi_i' = \psi_i + \alpha y_i h(x_i).$$

As in section 14.1.1, we treat ξ and α as distinct variables. To find values for them, we define two conditions, or equations, that they must satisfy, and then solve the equations simultaneously.

First, recall that in our intuitive description of the algorithm, h is used repeatedly for many subsequent time steps of BBM until its edge has been "used up." Thus, at the new time τ', and in the new chip positions ψ_i', it should be the case that h's weighted error is exactly $\frac{1}{2}$, so that its edge has been reduced to zero. This condition means that

$$\frac{\sum_{i=1}^{m} w(\psi_i', \tau') \mathbf{1}\{h(x_i) \neq y_i\}}{\sum_{i=1}^{m} w(\psi_i', \tau')} = \frac{1}{2},$$

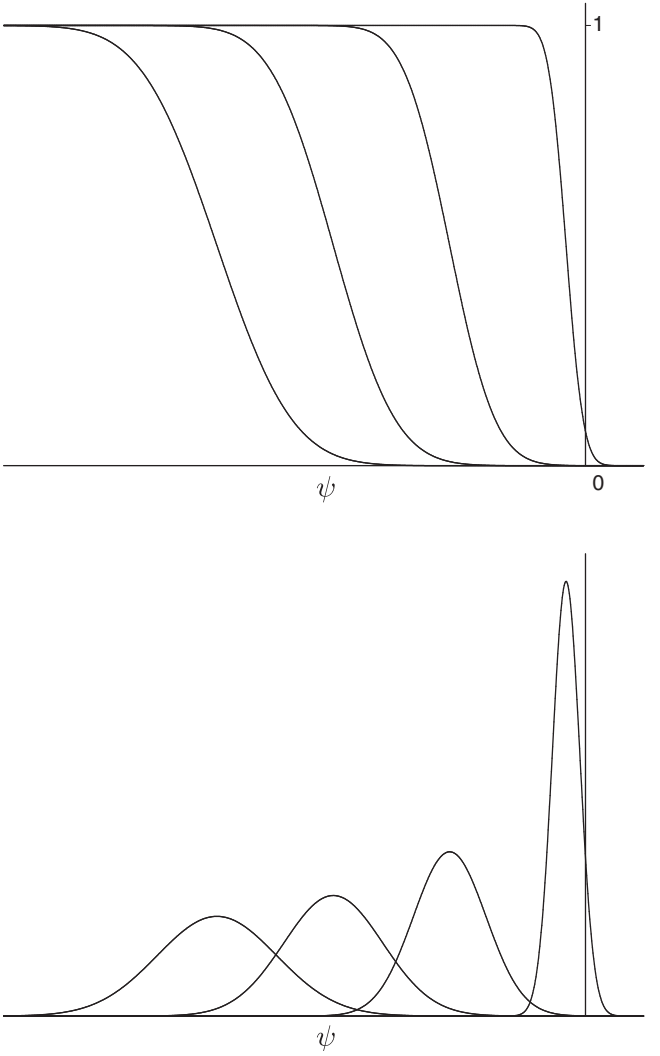

Figure 14.2
A plot of the potential function (top) and weighting function (bottom) when $\beta = 40$, as given in equations (14.9) and (14.15). In each figure, the curves are plotted, from left to right, with $\tau = 0.05$, 0.35, 0.65, and 0.95. (The four potential functions, although distinct, quickly become visually indistinguishable as they approach the limits of their range.) Based on the derivation in the text, the values for β and τ given here correspond to the variable settings for the plot of BBM's weighting function given in figure 13.4, which very closely resembles the smooth weighting function shown here.

which is equivalent to

$$\sum_{i=1}^{m} w(\psi_i',\ \tau')\ y_i h(x_i) = 0$$

or

$$\sum_{i=1}^{m} w(\psi_i + \alpha y_i h(x_i),\ \tau + \xi)\ y_i h(x_i) = 0. \tag{14.16}$$

This is the first equation that α and ξ should satisfy.

For the second condition, as discussed in section 14.1.1, we must continue to respect the key property that was used to analyze BBM: that the total potential of all the examples can never increase. In fact, in the continuous domain, if the chips and clock advance to a point at which the total potential has strictly *decreased*, then it turns out, by continuity of the potential function $\Phi(\psi,\ \tau)$, that it will always be possible to move the clock slightly further ahead while still ensuring that the total potential does not increase, relative to its starting value (see exercise 14.4). This means that here we can make an even stronger requirement, and insist that the total potential actually remain *unchanged*—neither increasing nor decreasing—so that

$$\sum_{i=1}^{m} \Phi(\psi_i,\ \tau) = \sum_{i=1}^{m} \Phi(\psi_i',\ \tau'),$$

or

$$\sum_{i=1}^{m} \Phi(\psi_i,\ \tau) = \sum_{i=1}^{m} \Phi(\psi_i + \alpha y_i h(x_i),\ \tau + \xi). \tag{14.17}$$

This is the second equation.

So α and ξ are chosen to satisfy equations (14.16) and (14.17), and then are used to update the clock and chip positions accordingly.

This entire process of finding weak hypotheses and solving for the appropriate updates to τ and the chip positions ψ_i repeats iteratively, and finally terminates when the clock τ reaches 1. Or, to avoid difficulties arising from the discontinuity in the potential function Φ when $\tau = 1$, we may wish to terminate when τ reaches some earlier cutoff $1 - c$, for some small $c > 0$. Upon termination, the final combined classifier is formed by taking a weighted majority vote of the weak hypotheses where each is assigned its associated weight α. The complete algorithm, called *BrownBoost* because of its connection to Brownian motion, is shown as algorithm 14.1. The procedure works in iterations which we index by r,

rather than t as in the rest of the book, to avoid confusion with the time steps of the BBM algorithm that conceptually is at its underpinning. We discuss the choice of β below.

14.2.2 Analysis

Because of the discontinuity in the potential function at $\tau = 1$, it is possible that no simultaneous solution will exist to BrownBoost's two equations (see exercise 14.6). However, if the algorithm is permitted to terminate when the clock τ reaches or exceeds $1 - c$, for some small $c > 0$, then the next theorem shows that a solution must always exist. (We do not discuss computational methods for actually finding a solution, but in practice, standard numerical methods can be applied.)

Theorem 14.1 Let Φ and w be defined as in equations (14.9) and (14.15). For any $\psi_1, \ldots, \psi_m \in \mathbb{R}$, $z_1, \ldots, z_m \in \{-1, +1\}$, $c > 0$, and $\tau \in [0, 1 - c)$, there exist $\alpha \in \mathbb{R}$ and $\tau' \in [\tau, 1]$ such that

$$\sum_{i=1}^{m} \Phi(\psi_i,\ \tau) = \sum_{i=1}^{m} \Phi(\psi_i + \alpha z_i,\ \tau'), \tag{14.18}$$

and either $\tau' \geq 1 - c$, or

$$\sum_{i=1}^{m} w(\psi_i + \alpha z_i,\ \tau')\ z_i = 0. \tag{14.19}$$

Proof We refer to a pair $\langle \alpha, \tau' \rangle$ with the properties stated in the theorem as a *BrownBoost solution*; our goal is to show that such a pair exists.

Let

$$\Pi(\alpha,\ \tau') \doteq \sum_{i=1}^{m} \Phi(\psi_i + \alpha z_i,\ \tau') \tag{14.20}$$

be the total potential of all chips after adjusting their positions by α, and after advancing the clock to τ'. In this notation, equation (14.18) holds if and only if

$$\Pi(0,\ \tau) = \Pi(\alpha,\ \tau').$$

Let

$$\mathcal{L} \doteq \left\{ \langle \alpha, \tau' \rangle : \Pi(\alpha,\ \tau') = \Pi(0,\ \tau), \alpha \in \mathbb{R}, \tau' \in [\tau, 1 - c] \right\} \tag{14.21}$$

be the *level set* of all pairs $\langle \alpha, \tau' \rangle$ satisfying equation (14.18), and with $\tau \leq \tau' \leq 1 - c$. To prove the theorem, it is sufficient (but not necessary) to find such a pair for which

Algorithm 14.1
The BrownBoost algorithm. The potential function $\Phi(\psi, \tau)$ and weighting function $w(\psi, \tau)$ are given in equations (14.9) and (14.15), respectively

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$
target error $\epsilon \in \left(0, \frac{1}{2}\right)$
clock cutoff $c \in [0, 1)$.

Initialize:

- Set β so that $\Phi(0,\ 0) = \epsilon$.
- Let $\tau_1 = 0$ and $\psi_{1,i} = 0$ for $i = 1, \ldots, m$.

For $r = 1, 2, \ldots$ until $\tau_r \geq 1 - c$:

- $D_r(i) = \dfrac{w(\psi_{r,i},\ \tau_r)}{\mathcal{Z}_r}$ for $i = 1, \ldots, m$,
  where $\mathcal{Z}_r$ is a normalization factor.
- Train weak learner using distribution D_r.
- Get weak hypothesis $h_r : \mathcal{X} \to \{-1, +1\}$.
- Aim: select h_r to minimalize the weighted error:

$$\mathbf{Pr}_{i \sim D_r}[h_r(x_i) \neq y_i].$$

- Find $\xi_r \geq 0$ and $\alpha_r \in \mathbb{R}$ such that $\tau_r + \xi_r \leq 1$,

$$\sum_{i=1}^{m} \Phi(\psi_{r,i},\ \tau_r) = \sum_{i=1}^{m} \Phi(\psi_{r,i} + \alpha_r y_i h_r(x_i),\ \tau_r + \xi_r),$$

  and either $\tau_r + \xi_r \geq 1 - c$ or

$$\sum_{i=1}^{m} w(\psi_{r,i} + \alpha_r y_i h_r(x_i),\ \tau_r + \xi_r)\ y_i h_r(x_i) = 0.$$

- Update:

$$\tau_{r+1} = \tau_r + \xi_r$$

$$\psi_{r+1,i} = \psi_{r,i} + \alpha_r y_i h_r(x_i) \quad \text{for } i = 1, \ldots, m.$$

Output the final hypothesis:

$$H(x) = \operatorname{sign}\left(\sum_{r=1}^{R} \alpha_r h_r(x)\right)$$

where R is the total number of iterations completed.

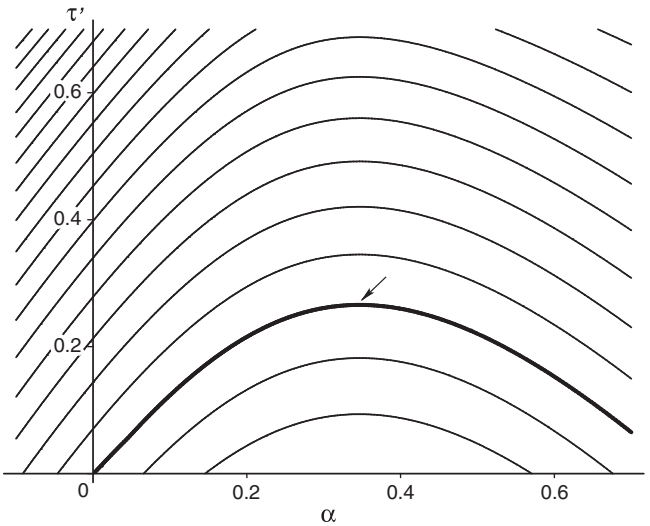

Figure 14.3
A typical contour plot for the function Π of equation (14.20), plotted on the very first round when $\tau = 0$ and $\psi_i = 0$ for all i. In this case, there are $m = 3$ training examples, one of which is misclassified (so $z_1 = z_2 = +1$ and $z_3 = -1$). The parameter β was chosen so that $\Phi(0,\ 0) = 1/4$. The level curves in the figure represent sets of points for which the value of Π is held constant. The level set $\mathcal{L}$ of interest (equation (14.21)) is the dark curve passing through $\langle 0, \tau \rangle$. The condition $\partial\Pi/\partial\alpha = 0$, which is the same as equation (14.19), is equivalent to the level curve becoming exactly horizontal; thus, in this case, a BrownBoost solution would exist at the very top of the dark curve, as indicated by the arrow.

either $\tau' = 1 - c$ or equation (14.19) holds, since these conditions imply that the pair is a BrownBoost solution. An example is shown in figure 14.3.

Note that, letting $\psi_i' = \psi_i + \alpha z_i$, we have by the chain rule from calculus that

$$\begin{aligned}\frac{\partial \Pi(\alpha,\ \tau')}{\partial \alpha} &= \sum_{i=1}^{m} \frac{\partial \Phi(\psi_i',\ \tau')}{\partial \psi_i'} \cdot \frac{d\psi_i'}{d\alpha} \\ &= -\sum_{i=1}^{m} w(\psi_i',\ \tau') z_i \qquad (14.22)\end{aligned}$$

by equation (14.15). Therefore, the left-hand side of equation (14.19), which is identical to the right-hand side of equation (14.22), always is equal to $-\partial\Pi/\partial\alpha$. So equation (14.19) is equivalent to the condition that $\partial\Pi/\partial\alpha = 0$.

In terms of a contour plot as in figure 14.3, this condition is equivalent to the level curve of interest becoming perfectly horizontal, as indicated in the figure. If this never happens, then intuitively the curve should eventually reach $\tau' = 1 - c$. In either case, we obtain the needed solution. Unfortunately, there are numerous potential complications; for instance,

in principle, the level set might not be connected, or could asymptote without either of the conditions above being satisfied.

To prove the theorem rigorously, we show first that if the α values of pairs in $\mathcal{L}$ are not bounded (so that they extend to $\pm\infty$), then a solution to equation (14.18) must exist at $\tau' = 1$, satisfying the theorem. Otherwise, when the α values are bounded, we argue that $\mathcal{L}$ is compact, and thus includes a pair with a maximal τ' value. Finally, we show that this pair is a BrownBoost solution.

Following this outline, suppose that the set of α-values occurring in $\mathcal{L}$, that is,

$$\mathcal{L}_1 \doteq \left\{\alpha : \langle \alpha, \tau' \rangle \in \mathcal{L} \text{ for some } \tau'\right\},$$

is unbounded. If

$$\sup \mathcal{L}_1 = \infty,$$

then there exists $\langle \alpha_1, \tau_1' \rangle, \langle \alpha_2, \tau_2' \rangle, \ldots$ such that $\langle \alpha_n, \tau_n' \rangle \in \mathcal{L}$ and $\alpha_n \to \infty$. This implies that as $\alpha_n \to \infty$,

$$\Phi(\psi_i + \alpha_n z_i,\ \tau_n') = \frac{1}{2}\text{erfc}\left(\frac{\psi_i + \alpha_n z_i + \beta(1-\tau_n')}{\sqrt{2\beta(1-\tau_n')}}\right)$$

is approaching 0 if $z_i = +1$, and 1 if $z_i = -1$, since the argument to the erfc is approaching $+\infty$ or $-\infty$, depending on z_i. This value of 0 or 1 is the same as $\Phi(\psi_i + \tilde{\alpha} z_i,\ 1)$ for some sufficiently large value of $\tilde{\alpha}$. Thus,

$$\Pi(0,\ \tau) = \Pi(\alpha_n,\ \tau_n') \to \Pi(\tilde{\alpha},\ 1).$$

It follows that the pair $\tilde{\alpha}$ and $\tilde{\tau}' = 1$ is a BrownBoost solution with $\Pi(0,\ \tau) = \Pi(\tilde{\alpha},\ \tilde{\tau}')$, and $\tilde{\tau}' \geq 1 - c$. (The case $\inf \mathcal{L}_1 = -\infty$ is handled symmetrically.)

Thus, we can assume henceforth that $\mathcal{L}_1$ is bounded and, therefore, that $\mathcal{L}$ is bounded as well. Furthermore, $\mathcal{L}$ is closed. For if $\langle \alpha_1, \tau_1' \rangle, \langle \alpha_2, \tau_2' \rangle, \ldots$ is a sequence of pairs in $\mathcal{L}$ converging to $\langle \hat{\alpha}, \hat{\tau}' \rangle$, then because Π is continuous on the region of interest,

$$\Pi(\alpha_n,\ \tau_n') \to \Pi(\hat{\alpha},\ \hat{\tau}').$$

Since the left-hand side is equal to the fixed value $\Pi(0,\ \tau)$ for all n, $\Pi(\hat{\alpha},\ \hat{\tau}')$ is as well. Further, since each τ_n' is in the closed set $[\tau, 1-c]$, $\hat{\tau}'$ must be also. Therefore, $\langle \hat{\alpha}, \hat{\tau}' \rangle \in \mathcal{L}$.

So $\mathcal{L}$ is compact, being both closed and bounded, and is also nonempty, since it includes $\langle 0, \tau \rangle$. These properties imply that there exists a pair $\langle \tilde{\alpha}, \tilde{\tau}' \rangle \in \mathcal{L}$ with maximum τ'-value, so that $\tilde{\tau}' \geq \tau'$ for all $\langle \alpha, \tau' \rangle \in \mathcal{L}$. We claim that $\langle \tilde{\alpha}, \tilde{\tau}' \rangle$ is the desired solution. Suppose, by way of reaching a contradiction, that it is not. Since it is in $\mathcal{L}$, this means that $\tilde{\tau}' < 1 - c$ and that $\partial\Pi/\partial\alpha$, evaluated at $\langle \tilde{\alpha}, \tilde{\tau}' \rangle$, is different from zero, by equation (14.22). Suppose $\partial\Pi/\partial\alpha$ is positive at this point (the argument when it is negative is symmetric). Then increasing $\tilde{\alpha}$ slightly causes Π to increase. That is, there exists $\varepsilon_+ > 0$ such that

$$\Pi(\tilde{\alpha} + \varepsilon_+,\ \tilde{\tau}') > \Pi(\tilde{\alpha},\ \tilde{\tau}') = \Pi(0,\ \tau). \tag{14.23}$$

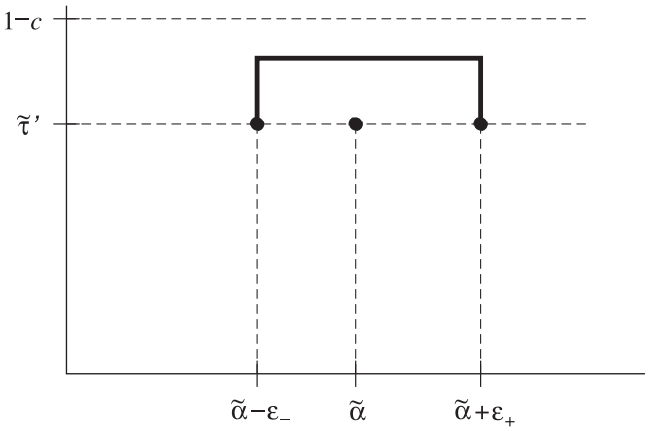

Figure 14.4
Construction of a path from $\langle \tilde{\alpha} - \varepsilon_-, \tilde{\tau}' \rangle$ to $\langle \tilde{\alpha} + \varepsilon_+, \tilde{\tau}' \rangle$ as used in the proof of theorem 14.1.

Likewise, there exists $\varepsilon_- > 0$ such that

$$\Pi(\tilde{\alpha} - \varepsilon_-,\ \tilde{\tau}') < \Pi(0,\ \tau). \tag{14.24}$$

We can now create a continuous path in the $\langle \alpha, \tau' \rangle$-plane from $\langle \tilde{\alpha} - \varepsilon_-, \tilde{\tau}' \rangle$ to $\langle \tilde{\alpha} + \varepsilon_+, \tilde{\tau}' \rangle$ in such a way that all of the points on the path, other than the endpoints, have τ' values smaller than $1 - c$ and strictly larger than $\tilde{\tau}'$. (See figure 14.4.) Because Π is continuous, equations (14.23) and (14.24) imply that there must be an intermediate point $\langle \hat{\alpha}, \hat{\tau}' \rangle$ on the path with $\Pi(\hat{\alpha},\ \hat{\tau}') = \Pi(0,\ \tau)$, that is, in the level set $\mathcal{L}$. However, $\hat{\tau}'$, being on the selected path, must be strictly larger than $\tilde{\tau}'$; this is a contradiction since $\tilde{\tau}'$ was itself chosen as the maximum among points in $\mathcal{L}$.

Thus, as claimed, $\langle \tilde{\alpha}, \tilde{\tau}' \rangle$ is the BrownBoost solution we seek, completing the proof. ■

There is no guarantee that BrownBoost's termination condition $\tau = 1$ (or even an earlier cutoff) will ever be reached. But suppose that it does terminate, and let us momentarily drop subscripts so that τ and ψ_i represent, respectively, the final time on the clock and the final position of chip i just before termination. If the algorithm halts with $\tau = 1$, then the training error of its final hypothesis H is simple to analyze using the same idea as in corollary 13.4: At time $\tau = 1$, the average potential

$$\frac{1}{m} \sum_{i=1}^{m} \Phi(\psi_i,\ 1)$$

is exactly equal to the training error by equation (14.14) and by H's definition. Since this average potential never changed throughout the algorithm's execution, it must be equal to the average initial potential, which is

$$\Phi(0,\ 0) = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\beta}{2}}\right). \tag{14.25}$$

Thus, the algorithm takes as input a parameter $\epsilon > 0$, which is the target error, and sets β so that equation (14.25) will be equal to ϵ. Note that this ensures that the final error at $\tau = 1$ will be *exactly* ϵ.

If, as discussed in section 14.1.3, $\Phi(0,\ 1)$ is instead defined to be $\frac{1}{2}$, we still obtain an exact result for the training error at $\tau = 1$, but slightly redefined so that a prediction of 0 counts as only half a mistake. This alternative definition is reasonable since such a prediction can be regarded as a random guess that is correct with probability exactly $\frac{1}{2}$.

If the algorithm is permitted to terminate at some time $\tau < 1$, we can also obtain training error bounds. If the final hypothesis H makes a mistake on some training example i, so that $\psi_i \leq 0$, then because $\Phi(\psi,\ \tau)$ is decreasing in ψ (since the erfc function is decreasing), we must have $\Phi(\psi_i,\ \tau) \geq \Phi(0,\ \tau)$. Since Φ is never negative, we thus have, in general,

$$\Phi(0,\ \tau) \cdot \mathbf{1}\{\psi_i \leq 0\} \leq \Phi(\psi_i,\ \tau).$$

Averaging both sides over all examples gives

$$\Phi(0,\ \tau) \cdot \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\psi_i \leq 0\} \leq \frac{1}{m}\sum_{i=1}^{m} \Phi(\psi_i,\ \tau).$$

The left-hand side is $\Phi(0,\ \tau)$ times the training error. The right-hand side is the average potential at time τ, which, as noted above, is equal to the initial potential $\Phi(0,\ 0) = \epsilon$. Therefore, using Φ's definition in equation (14.9), if $\tau \geq 1 - c$, then the training error is at most

$$\frac{2\epsilon}{\mathrm{erfc}\left(\sqrt{\beta c/2}\right)}. \tag{14.26}$$

Since the denominator approaches 1 as $c \to 0$, this bound can be made arbitrarily close to 2ϵ by choosing c sufficiently small. (For a technique that instead approaches ϵ, see exercise 14.7.)

14.3 AdaBoost as a Special Case of BrownBoost

In section 13.3.2, we saw that the behavior of BBM on the initial rounds converges to that of NonAdaBoost, a nonadaptive version of AdaBoost, as T becomes large. Correspondingly, in this section, we will see that AdaBoost, in its usual adaptive form, can be derived from BrownBoost, an adaptive version of BBM, by taking an appropriate limit, namely, the limit as the error parameter ϵ is taken to zero. Thus, AdaBoost can be viewed as a special case of BrownBoost in which a final training error of zero is anticipated. Or, turning this statement

around, BrownBoost can be viewed as a generalization of AdaBoost in which a positive final training error is expected.

To show this, we need to argue that, under this limit, the distribution over examples computed by BrownBoost on every round converges to that for AdaBoost, and that the weighted-majority combined classifiers are the same for the two algorithms. We presume here that the same sequence of weak classifiers is returned by the weak learning algorithm for either boosting algorithm, a reasonable assumption since the distributions on which they are trained will be nearly the same; however, as pointed out in section 13.3.2, the assumption is unprovable in general for reasons of numerical instability. Thus, under this assumption, to show that the combined hypotheses are the same, it suffices to show that the coefficients α_r computed by the two algorithms are the same, in the limit.

Although we are interested in what happens when $\epsilon \to 0$, we will find it more convenient in what follows to frame the discussion in terms of the limit $\beta \to \infty$. Because of how BrownBoost chooses β as a function of ϵ (so that equation (14.25) is equal to ϵ), these two cases are exactly equivalent.

To clarify the notation, let us write $\psi_{r,i}^{\beta}$, D_r^{β}, α_r^{β}, etc. for the variables used by BrownBoost (algorithm 14.1) when run with parameter β (for some corresponding choice of ϵ), and let D_r^*, α_r^*, etc. denote the variables used by AdaBoost (algorithm 1.1 (p. 5)), where we now use r rather than t to index the round number for both algorithms. For BrownBoost, we assume that the clock cutoff has been fixed to an arbitrary positive constant $c \in (0, 1)$. Further, we assume that BrownBoost is run *lingeringly*, meaning that it never halts unless forced to do so; that is, on every round r, it chooses a solution to its equations with $\tau_{r+1} < 1 - c$, unless no such solution exists.

In precise terms, we prove the following concerning BrownBoost's behavior under the limit $\beta \to \infty$ (which, as just noted, is equivalent to $\epsilon \to 0$).

Theorem 14.2 Suppose BrownBoost (with fixed cutoff $c \in (0, 1)$) and AdaBoost are run on the same dataset and are provided with the same round-by-round sequence of weak hypotheses $h_1, h_2, \ldots$, none of which are perfectly accurate or perfectly inaccurate on the entire dataset. Assume the notation above, and that BrownBoost is run lingeringly. Let r be any positive integer. Then for all sufficiently large β, BrownBoost will not halt before reaching round r. Furthermore, as $\beta \to \infty$, BrownBoost's distribution and hypothesis weights on round r converge to those for AdaBoost; that is,

$$D_r^{\beta} \to D_r^* \tag{14.27}$$

and

$$\alpha_r^{\beta} \to \alpha_r^*. \tag{14.28}$$

Note that, because we assume the same sequence of weak classifiers $h_1, h_2, \ldots$ is received by both algorithms, equation (14.28) implies that

$$\sum_{r=1}^{R} \alpha_r^{\beta} h_r(x) \to \sum_{r=1}^{R} \alpha_r^{*} h_r(x)$$

for any R and x, so that the weighted-majority combined classifiers will also be the same for AdaBoost and BrownBoost in the limit (except possibly in the degenerate case that the sum on the right is exactly zero).

Proof Assume inductively that equation (14.28) holds on rounds $1, \ldots, r-1$. We wish to show that equations (14.27) and (14.28) hold on the current round r. We also must show that the time τ does not reach the cutoff of $1-c$ within r rounds. Thus, for β sufficiently large, we assume inductively that

$$\tau_r^{\beta} < 1 - c, \tag{14.29}$$

and will show that the same holds for τ_{r+1}^{β}, ensuring that the algorithm does not terminate.

Let us fix r, and drop it from the notation when it is clear from context so that $\tau^{\beta} = \tau_r^{\beta}$, $\psi_i^{\beta} = \psi_{r,i}^{\beta}$, and so on. We also define $\psi_i^{*} = \psi_{r,i}^{*}$ to be the unnormalized margin computed by AdaBoost:

$$\psi_i^{*} \doteq y_i \sum_{r'=1}^{r-1} \alpha_{r'}^{*} h_{r'}(x_i).$$

Note that by our inductive hypothesis on equation (14.28),

$$\psi_i^{\beta} \to \psi_i^{*} \tag{14.30}$$

as $\beta \to \infty$ since

$$\psi_i^{\beta} = y_i \sum_{r'=1}^{r-1} \alpha_{r'}^{\beta} h_{r'}(x_i).$$

Finally, we will sometimes write w^{β} and Φ^{β} to make explicit the dependence of BrownBoost's weighting and potential functions on the parameter β.

So, to summarize, given our inductive assumptions, we need to prove equations (14.27) and (14.28), and that BrownBoost does not halt. We prove each of these three in turn.

Lemma 14.3 Under the assumptions and notation above,

$$D_r^{\beta} \to D_r^{*}.$$

Proof We can rewrite the weighting function of equation (14.15) as follows:

$$w^{\beta}(\psi,\ \tau) \propto \exp\left(-\frac{\psi^2 + 2\psi\beta(1-\tau) + \beta^2(1-\tau)^2}{2\beta(1-\tau)}\right)$$

$$= \exp\left(-\frac{\psi^2}{2\beta(1-\tau)} - \psi - \frac{\beta(1-\tau)}{2}\right)$$
$$\propto \exp\left(-\frac{\psi^2}{2\beta(1-\tau)} - \psi\right)$$
$$= \exp\left(-\psi\left(1 + \frac{\psi}{2\beta(1-\tau)}\right)\right) \tag{14.31}$$

(where $f \propto g$ means f is equal to g times a positive factor that does not depend on ψ). Thus,

$$w^\beta(\psi_i^\beta,\ \tau^\beta) \propto \exp\left(-\psi_i^\beta\left(1 + \frac{\psi_i^\beta}{2\beta(1-\tau^\beta)}\right)\right). \tag{14.32}$$

By equations (14.29) and (14.30), it follows that the expression on the right converges to $\exp(-\psi_i^*)$, which is exactly proportional to $D_r^*(i)$, the normalized weight assigned by AdaBoost to training example i (see, for instance, equation (3.2)). Since $D_r^\beta(i)$ is proportional to equation (14.32), equation (14.27) follows immediately. ∎

Let $z_i \doteq y_i h(x_i)$. As in the proof of theorem 14.1, let

$$\Pi^\beta(\alpha,\ \tau') \doteq \sum_{i=1}^m \Phi^\beta(\psi_i^\beta + \alpha z_i,\ \tau')$$

be the total potential of all the training examples after adjusting their positions by α and advancing the clock to τ'. At the solution $\langle \alpha^\beta, \tau'^\beta \rangle$ found by BrownBoost, where $\tau'^\beta \doteq \tau_{r+1}^\beta$, we will have

$$\Pi^\beta(\alpha^\beta,\ \tau'^\beta) = \Pi^\beta(0,\ \tau^\beta). \tag{14.33}$$

Further, the solution either will have $\tau'^\beta \geq 1 - c$, or will satisfy

$$\sum_{i=1}^m w^\beta(\psi_i^\beta + \alpha^\beta z_i,\ \tau'^\beta)\ z_i = 0. \tag{14.34}$$

To show that BrownBoost does not halt on the current round, we apply theorem 14.1, where we proved that such a solution must exist, and we show that, for β large, the cutoff $1 - c$ cannot be attained by the solution guaranteed by this theorem.

Lemma 14.4 Under the assumptions and notation above,

$\tau'^\beta < 1 - c$.

Proof The proof of theorem 14.1 shows specifically that a solution to BrownBoost's equations will exist with $\tau'^{\beta} = 1$ or $\tau'^{\beta} \leq 1 - c$. To prove the lemma, we show that the cases $\tau'^{\beta} = 1$ and $\tau'^{\beta} = 1 - c$ are not possible, for β large, so that a solution with $\tau'^{\beta} < 1 - c$ must necessarily exist and be chosen by BrownBoost, which we assume is being run lingeringly.

First, note that $\Phi^{\beta}(\psi,\ 1)$ is always in $\{0, 1\}$, so if $\tau'^{\beta} = 1$, then $\Pi^{\beta}(\alpha^{\beta},\ \tau'^{\beta})$ must be an integer. But because the potential remains constant throughout the execution of BrownBoost,

$$\Pi^{\beta}(\alpha^{\beta},\ \tau'^{\beta}) = m \cdot \Phi^{\beta}(0,\ 0) = m\epsilon, \tag{14.35}$$

which is not an integer for β large and ϵ correspondingly small but positive. Thus, $\tau'^{\beta} \neq 1$. (This argument assumes $\Phi(0,\ 1) \doteq 1$, but can be straightforwardly modified if instead $\Phi(0,\ 1) \doteq \frac{1}{2}$.)

Suppose next that $\tau'^{\beta} = 1 - c$. We show that this leads to a contradiction, for β large. Let b be any constant for which $c < b < 1$, and let

$$d \doteq \sqrt{bc} - c > 0. \tag{14.36}$$

Since the current weak hypothesis h is neither perfectly accurate nor perfectly inaccurate, there must exist i and i' for which $z_i = -z_{i'}$. Further, for β sufficiently large,

$$\psi_i^{\beta} + \psi_{i'}^{\beta} \leq 2\beta d$$

since, by equation (14.30), the left-hand side is converging to a fixed value while the right-hand side is growing to infinity. Because $z_i + z_{i'} = 0$, this implies that

$$(\psi_i^{\beta} + \alpha^{\beta} z_i) + (\psi_{i'}^{\beta} + \alpha^{\beta} z_{i'}) \leq 2\beta d,$$

which means at least one of the parenthesized expressions on the left, say the first, is at most βd, that is,

$$\psi_i^{\beta} + \alpha^{\beta} z_i \leq \beta d.$$

Rewriting, using equation (14.36), this gives

$$\frac{\psi_i^{\beta} + \alpha^{\beta} z_i + \beta c}{\sqrt{2\beta c}} \leq \sqrt{\frac{\beta b}{2}}.$$

Thus,

$$\begin{aligned}
\Pi^{\beta}(\alpha^{\beta},\ \tau'^{\beta}) &\geq \Phi^{\beta}(\psi_i^{\beta} + \alpha^{\beta} z_i,\ 1 - c) \\
&= \frac{1}{2}\mathrm{erfc}\left(\frac{\psi_i^{\beta} + \alpha^{\beta} z_i + \beta c}{\sqrt{2\beta c}}\right) \\
&\geq \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\beta b}{2}}\right)
\end{aligned} \tag{14.37}$$

by Φ's definition in equation (14.9), and because the erfc function is decreasing. Using a standard approximation to the erfc function,

$$\frac{2}{\sqrt{\pi}} \cdot \frac{e^{-u^2}}{u+\sqrt{u^2+2}} \leq \operatorname{erfc}(u) \leq \frac{2}{\sqrt{\pi}} \cdot \frac{e^{-u^2}}{u+\sqrt{u^2+4/\pi}}, \tag{14.38}$$

which holds for all $u > 0$, it follows from equation (14.37) that

$$\Pi^{\beta}(\alpha^{\beta},\ \tau'^{\beta}) \geq \exp\left(-\beta\left(\frac{b}{2}+o(1)\right)\right) \tag{14.39}$$

(where $o(1)$ represents a quantity that approaches zero as $\beta \to \infty$). On the other hand, as argued earlier,

$$\Pi^{\beta}(\alpha^{\beta},\ \tau'^{\beta}) = m \cdot \Phi^{\beta}(0,\ 0) = \frac{m}{2}\operatorname{erfc}\left(\sqrt{\frac{\beta}{2}}\right) \leq \exp\left(-\beta\left(\frac{1}{2}-o(1)\right)\right) \tag{14.40}$$

where we have again applied equation (14.38). Since $b < 1$, when β is large, equations (14.39) and (14.40) are in contradiction. ∎

It remains only to show that $\alpha^{\beta} \to \alpha^*$. In very rough terms, this can be seen as follows: By lemma 14.4, $\tau'^{\beta} < 1-c$, and therefore equation (14.34) must hold at the solution $\langle \alpha^{\beta}, \tau'^{\beta} \rangle$. Approximating $w^{\beta}(\psi,\ \tau)$ by $\exp(-\psi)$ based on the proof of lemma 14.3, this equation becomes

$$\sum_{i=1}^{m} \exp\left(-(\psi_i^{\beta} + \alpha^{\beta} z_i)\right)\ z_i = 0. \tag{14.41}$$

Recall from section 7.1 that α^* is chosen by AdaBoost to minimize

$$\sum_{i=1}^{m} \exp\left(-(\psi_i^{*} + \alpha^{*} z_i)\right),$$

in other words, to have derivative with respect to α^* equal to zero:

$$\sum_{i=1}^{m} \exp\left(-(\psi_i^{*} + \alpha^{*} z_i)\right)\ z_i = 0. \tag{14.42}$$

Since $\psi_i^{\beta} \to \psi_i^*$, the matching equations (14.41) and (14.42) imply that $\alpha^{\beta} \to \alpha^*$. The next lemma provides a more rigorous proof of equation (14.28) based on this idea.

Lemma 14.5 Let $\delta > 0$. Under the assumptions and notation above, for β sufficiently large, $|\alpha^{\beta} - \alpha^*| < \delta$.

Proof For all i, we must have

$$\psi_i^\beta + \alpha^\beta z_i + \beta(1-\tau'^\beta) > 0 \tag{14.43}$$

for β large. Otherwise, if this were not the case for some i, then

$$\begin{aligned}\Pi^\beta(\alpha^\beta,\ \tau'^\beta) &\geq \Phi^\beta(\psi_i^\beta + \alpha^\beta z_i,\ \tau'^\beta)\\ &= \frac{1}{2}\mathrm{erfc}\left(\frac{\psi_i^\beta + \alpha^\beta z_i + \beta(1-\tau'^\beta)}{\sqrt{2\beta(1-\tau'^\beta)}}\right)\\ &\geq \tfrac{1}{2}\mathrm{erfc}(0) = \tfrac{1}{2}\end{aligned}$$

by Φ's definition in equation (14.9). This contradicts equation (14.35) for β large (and ϵ therefore small). Thus, equation (14.43) holds for all i, which is equivalent to saying that

$$\alpha^\beta \in (M_-^\beta, M_+^\beta)$$

where

$$\begin{aligned}M_-^\beta &\doteq \max_{i:z_i=+1}\left[-\psi_i^\beta - \beta(1-\tau'^\beta)\right]\\ M_+^\beta &\doteq \min_{i:z_i=-1}\left[\psi_i^\beta + \beta(1-\tau'^\beta)\right].\end{aligned} \tag{14.44}$$

Let

$$W^\beta(\alpha,\ \tau') \doteq \sum_{i=1}^m \exp\left(-(\psi_i^\beta + \alpha z_i)\left(1 + \frac{\psi_i^\beta + \alpha z_i}{2\beta(1-\tau')}\right)\right) z_i.$$

By equations (14.31) and (14.22),

$$W^\beta(\alpha,\ \tau') \propto \sum_{i=1}^m w^\beta(\psi_i^\beta + \alpha z_i,\ \tau')\, z_i = -\frac{\partial \Pi^\beta(\alpha,\ \tau')}{\partial \alpha}. \tag{14.45}$$

Therefore, equation (14.34), which must be satisfied at the solution $\langle \alpha^\beta, \tau'^\beta \rangle$, is equivalent to the condition

$$W^\beta(\alpha^\beta,\ \tau'^\beta) = 0. \tag{14.46}$$

Further, for β sufficiently large, we claim that $W^\beta(\alpha,\ \tau')$ is decreasing in α for $\alpha \in (M_-^\beta, M_+^\beta)$. To see this, note first that $\mathrm{erfc}(u)$ is convex for $u > 0$, as can be seen in figure 14.1. Thus, $\Phi^\beta(\psi_i^\beta + \alpha z_i,\ \tau')$—which, by its definition in equation (14.9), is equal to erfc evaluated at a linear function of α—is convex in α for α satisfying equation (14.43). In turn, this implies that $\Pi^\beta(\alpha,\ \tau')$, being the sum of several convex functions, is also convex

in α, for $\alpha \in (M_-^\beta, M_+^\beta)$. Therefore, $\partial \Pi^\beta(\alpha, \tau')/\partial\alpha$ is increasing in α, which means, by equation (14.45), that $W^\beta(\alpha, \tau')$ is decreasing in α, for α in this interval.

For any α and $\tau < 1$, it is clear, using equation (14.30), that

$$W^\beta(\alpha, \tau) \to W^*(\alpha)$$

as $\beta \to \infty$, where

$$W^*(\alpha) \doteq \sum_{i=1}^{m} \exp\left(-(\psi_i^* + \alpha z_i)\right) z_i,$$

the corresponding function for AdaBoost. Moreover, this convergence is uniform for $\tau \in [0, 1-c]$, meaning that the convergence happens simultaneously for all values of τ so that

$$\sup_{0 \le \tau \le 1-c} \left| W^\beta(\alpha, \tau) - W^*(\alpha) \right| \to 0.$$

From their definition in equation (14.44), it can be seen that $M_-^\beta \to -\infty$ and $M_+^\beta \to +\infty$ as $\beta \to \infty$, by equation (14.30), and since $0 \le \tau'^\beta < 1 - c$. Thus, $\alpha^* \in (M_-^\beta, M_+^\beta)$ for β sufficiently large.

It also can be checked that W^* is strictly decreasing and, by equation (14.42), is equal to zero at α^*. This implies that $W^*(\alpha^* + \delta) < 0$, so for β sufficiently large,

$$W^\beta(\alpha^* + \delta, \tau') < 0$$

for all $\tau' \in [0, 1-c]$. Since $W^\beta(\alpha, \tau'^\beta)$ is decreasing in α for $\alpha \in (M_-^\beta, M_+^\beta)$, and since $\alpha^* > M_-^\beta$, it follows that $W^\beta(\alpha, \tau'^\beta) < 0$ for $\alpha^* + \delta \le \alpha < M_+^\beta$, precluding a solution to equation (14.46) in this interval. And we have already argued that the solution cannot happen for any $\alpha \ge M_+^\beta$. Thus, we have eliminated all possibilities for the solution α^β to be at least $\alpha^* + \delta$.

Therefore, $\alpha^\beta < \alpha^* + \delta$. A similar argument shows that $\alpha^\beta > \alpha^* - \delta$, completing the proof. ∎

Thus, we have also completed the proof of theorem 14.2, having shown that as $\beta \to \infty$, which is the same as $\epsilon \to 0$, the behavior of BrownBoost converges exactly to that of AdaBoost for any finite number of rounds. ∎

14.4 Experiments with Noisy Data

Theorem 14.2 suggests AdaBoost may be best matched with a setting in which the training error can be driven to zero. This agrees with the training-error analysis of section 3.1, where we saw how the weak learning assumption suffices to assure perfect training accuracy in a very small number of rounds. But this view is also consistent with AdaBoost's susceptibility

to noise, discussed in section 12.3.3, and its general propensity to direct inordinate attention to the hardest examples, which might well have been corrupted or mislabeled.

BrownBoost, on the other hand, may have a better chance of handling such noisy settings. First, the algorithm explicitly anticipates a nonzero training error of $\epsilon > 0$, as seen in section 14.2.2. And furthermore, as was the case for BBM, as discussed in section 13.3.3, BrownBoost's weighting function causes it to deliberately "give up" on the hardest examples, focusing instead on those examples that still have a reasonable chance of eventually being correctly classified.

As an illustration of the improvement in performance that might be possible, BrownBoost was compared experimentally with AdaBoost on the noisy, synthetic learning problem described in section 12.3.2, which we showed in that section will ultimately cause AdaBoost to perform very poorly under appropriate limits. As earlier explained, examples in this setting are binary vectors of length $N = 2n + 11$ with weak hypotheses identified with individual coordinates; here, the cases $n = 5$ and $n = 20$ were tested. The "clean" label associated with each example can be computed as a simple majority vote over a subset of its coordinates. The actual observed labels, however, are noisy versions of the clean labels which have been corrupted (that is, negated) with a noise rate of η; noise rates of 0% (no noise), 5%, and 20% were considered in the experiments.

BrownBoost was run with various values of ϵ, and the one giving lowest training error was selected for use during testing. AdaBoost.L, the version of AdaBoost based on logistic loss from section 7.5.2, was also compared, since its more moderate weighting function suggests that it might handle noise and outliers better than AdaBoost. (Note, however, that AdaBoost.L must also eventually perform very poorly on this data, by arguments similar to those in section 12.3.2; see also exercise 12.9.)

Training sets of $m = 1000$ and 10,000 examples were used. Each algorithm was run for a maximum of 1000 rounds but, as in algorithm 14.1, BrownBoost can stop early if the clock τ reaches $1 - c$, where a cutoff of $c = 0.01$ was used throughout.

Table 14.2 reports the error for each algorithm on a separate test set of 5000 *uncorrupted* examples, that is, with labels that are clean. Consistent with what was proved in section 12.3.2, AdaBoost does quite poorly on this problem. AdaBoost.L does better in the easiest case that $n = 5$ and $\eta = 5\%$, but otherwise performs almost as badly as AdaBoost. (Observe, incidentally, that when $n = 20$, performance actually gets *worse* for both algorithms when given *more* data.) BrownBoost, on the other hand, performs very well, attaining almost perfect test accuracy in most cases when $n = 5$, and giving far better accuracy than either AdaBoost or AdaBoost.L in the harder case that $n = 20$.

These same algorithms were also tested on real-world, benchmark datasets, artificially corrupted with additional label noise at rates of 0%, 10% and 20%. Here, each boosting method was combined with the alternating decision tree algorithm of section 9.4. Also, a variant of BrownBoost was used in which the "boundary condition" of equation (14.14) is replaced by

Table 14.2
The results of running AdaBoost, AdaBoost.L, and BrownBoost on the noisy, synthetic learning problem of section 12.3.2 with various settings of n, m, and η

		$m = 1000$			$m = 10{,}000$		
n	η	AdaBoost	AdaBoost.L	BrownBoost	AdaBoost	AdaBoost.L	BrownBoost
5	0%	0.0	0.0	0.0	0.0	0.0	0.0
	5%	19.4	2.7	0.4	8.5	0.0	0.0
	20%	23.1	22.0	2.2	21.0	17.4	0.0
20	0%	0.0	3.7	0.8	0.0	0.0	0.1
	5%	31.1	29.9	10.7	41.3	36.8	5.4
	20%	30.4	30.2	21.1	36.9	36.1	12.0

Each entry shows percent error on *clean* (uncorrupted) test examples. All results are averaged over ten random repetitions of the experiment.

Table 14.3
The results of running AdaBoost, AdaBoost.L, and BrownBoost on the "letter" and "satimage" benchmark datasets

Dataset	η	AdaBoost	AdaBoost.L	BrownBoost
letter	0%	3.7	3.7	4.2
	10%	10.8	9.4	7.0
	20%	15.7	13.9	10.5
satimage	0%	4.9	5.0	5.2
	10%	12.1	11.9	6.2
	20%	21.3	20.9	7.4

After converting to binary by combining the classes into two arbitrary groups, each dataset was split randomly into training and test sets, and corrupted for training with artificial noise at rate η. The entries of the table show percent error on *uncorrupted* test examples. All results are averaged over 50 random repetitions of the experiment.

$$\Phi(\psi,\ 1) = \mathbf{1}\{\psi \leq \vartheta\}\,, \tag{14.47}$$

for some parameter $\vartheta \geq 0$; see exercise 14.2. Both ϵ and ϑ were chosen by training on 75% of the training data using various settings of these parameters, and then choosing the single setting that performed best on the remaining, held-out training examples.

Table 14.3 shows percent error on clean, uncorrupted test examples. Again, BrownBoost performs much better than the other algorithms in the presence of noise.

Summary

In this chapter, we have described a technique for making BBM adaptive by porting it to a continuous-time setting. We saw that BrownBoost, the resulting algorithm, is a

generalization of AdaBoost, but one which may have favorable properties in its handling of noisy data and outliers.

Bibliographic Notes

The results of sections 14.1, 14.2, and 14.3 are an elaboration and extension of the work of Freund [89] on the original version of BrownBoost, as well as later work by Freund and Opper [92] which connected the continuous-time framework with drifting games [201], and also introduced an approach based on differential equations similar to that given in section 14.1.3.

The experiments summarized in section 14.4 were conducted jointly with Evan Ettinger and Sunsern Cheamanunkul.

Further background on the central limit theorem and the convergence of distributions can be found in any standard text on probability, such as [21, 33, 84, 215]. More about Brownian motion and stochastic differential equations can be found, for instance, in [61, 131, 177, 216]. Equation (14.38) appears in Gautschi [105].

Some of the exercises in this chapter are based on material from [92].

Exercises

14.1 Let $\Phi(\psi,\ \tau)$ be defined as in equation (14.9), for some $\beta > 0$.

a. Verify that the partial differential equation given in equation (14.13) is satisfied for all $\psi \in \mathbb{R}$ and $\tau \in [0, 1)$.

b. Verify that the boundary condition given in equation (14.14) is satisfied away from $\psi = 0$. More specifically, let $\langle \psi_n, \tau_n \rangle$ be any sequence of pairs in $\mathbb{R} \times [0, 1)$. Show that if $\langle \psi_n, \tau_n \rangle \to \langle \psi, 1 \rangle$ as $n \to \infty$, where $\psi \neq 0$, then $\Phi(\psi_n,\ \tau_n) \to \mathbf{1}\{\psi \leq 0\}$.

c. For all $v \in [0, 1]$, show there exists a sequence $\langle \psi_n, \tau_n \rangle$ in $\mathbb{R} \times [0, 1)$ such that $\langle \psi_n, \tau_n \rangle \to \langle 0, 1 \rangle$ and $\Phi(\psi_n,\ \tau_n) \to v$ as $n \to \infty$.

14.2 Let $\vartheta > 0$ be a fixed value representing a desired margin. Suppose equation (14.14) is replaced with the modified boundary condition given in equation (14.47).

a. Find an expression for $\Phi(\psi,\ \tau)$ which satisfies equations (14.13) and (14.47) in the sense of exercise 14.1.

b. Find an expression for the weighting function $w(\psi,\ \tau)$ that corresponds to this modified potential function.

Note that if BrownBoost is used with these modified versions of Φ and w (for given values of $\epsilon > 0$ and $\vartheta > 0$), and if the algorithm stops at time $\tau = 1$, then the fraction of training examples i with margin $\psi_i \leq \vartheta$ will be exactly ϵ.

c. Show how this potential function could alternatively be derived in the limit $T \to \infty$ from the potential associated with the version of BBM given in exercise 13.5 (for an appropriate choice of θ in terms of T, β, and ϑ).

14.3 Suppose that $\Phi_t(s)$ and $w_t(s)$ are redefined as in exercise 13.6, with α hardwired using the value derived in part (c) of that exercise. We saw earlier that these choices, in BBM, lead to NonAdaBoost. Here, we explore what happens in the continuous-time limit.

a. For fixed $\beta > 0$, $\psi \in \mathbb{R}$, and $\tau \in [0, 1]$, let s, t, and γ be chosen, as functions of T, to satisfy equations (14.1), (14.2), and (14.7) (or to satisfy them as nearly as possible, subject to s and t being integers). Compute $\Phi(\psi, \tau)$, the limit of $\Phi_t(s)$ as $T \to \infty$. Also, use equation (14.15) to compute $w(\psi, \tau)$. Your final answers should be in terms of β, ψ, and τ only. [*Hint:* Use the fact that for any $a \in \mathbb{R}$, $\lim_{x\to\infty}(1 + a/x)^x = e^a$.]

b. Explain why we expect that your answer in part (a) for $\Phi(\psi, \tau)$ should satisfy equation (14.13). Then verify that it does.

In the remainder of this exercise, we consider a variant of BrownBoost (algorithm 14.1) in which the potential and weighting functions have been replaced by those in part (a). We use a cutoff of $c = 0$, and assume $\beta > 0$ throughout.

c. For this modified version of BrownBoost, show that there always exists a solution to the algorithm's two main equations. That is, for any $\psi_1, \ldots, \psi_m \in \mathbb{R}$, $z_1, \ldots, z_m \in \{-1, +1\}$, and $\tau \in [0, 1)$, prove that there exist $\alpha \in \mathbb{R}$ and $\tau' \in [\tau, 1]$ such that equation (14.18) holds, and either $\tau' = 1$ or equation (14.19) holds (using the revised definition of Φ, of course). You can assume that the z_i's are not all the same sign. Also show that this solution is unique, except possibly when $\tau' = 1$.

d. Suppose, for some integer $R > 0$, that $\tau_{R+1} < 1$ (so that the clock has not run out within R rounds). Show that modified BrownBoost's behavior on these first R rounds is identical to that of AdaBoost (algorithm 1.1). That is, assuming the same sequence of weak hypotheses $h_1, \ldots, h_R$ is provided to both algorithms, prove that $D_r^{\text{AB}} = D_r^{\text{BB}}$ and $\alpha_r^{\text{AB}} = \alpha_r^{\text{BB}}$ for $r = 1, \ldots, R$, where we use superscripts AB and BB to distinguish the variables of AdaBoost and (modified) BrownBoost, respectively (and where we use r instead of t to denote round number).

e. Let $R > 0$ be a fixed integer. Show that for β sufficiently large, $\tau_{R+1} < 1$. (You can assume that the sequence of weak hypotheses is fixed and independent of β.)

f. Given $\epsilon \in \left(0, \frac{1}{2}\right)$, explain how a stopping criterion could be added to AdaBoost which would be equivalent to the stopping criterion $\tau_r = 1$ used in modified BrownBoost. If the empirical γ-weak learning assumption holds, for some $\gamma > 0$, must (modified) BrownBoost necessarily halt within a finite number of rounds? Why or why not?

Exercises 14.4, 14.5, and 14.6 explore the nature of BrownBoost solutions in greater detail. For all of these, we adopt the setup and notation of theorem 14.1, including the definition of $\Pi(\alpha,\ \tau')$ given in equation (14.20). Also, we let $\epsilon \doteq \Pi(0,\ \tau)/m$, and we assume $\epsilon \in \left(0,\ \frac{1}{2}\right)$, and that $\beta > 0$ is given and fixed.

14.4 Suppose there exist $\alpha \in \mathbb{R}$ and $\tau' \in [\tau, 1-c)$ which satisfy equation (14.19), but for which $\Pi(\alpha,\ \tau') < \Pi(0,\ \tau)$. Show that there exists a BrownBoost solution $\langle \tilde{\alpha}, \tilde{\tau}' \rangle$ with $\tilde{\tau}' > \tau'$.

14.5 Let $\delta \doteq \frac{1}{2} - \frac{1}{2m}\sum_{i=1}^{m} z_i$, and assume $\delta \in \left(0,\ \frac{1}{2}\right)$.

a. Show that if ϵ is not an integer multiple of $1/m$, then there cannot exist a BrownBoost solution $\langle \alpha, \tau' \rangle$ with $\tau' = 1$. Also, show that if $\epsilon \neq \delta$, then there must exist a BrownBoost solution with $\tau' \leq 1 - c$.

b. Consider the function

$$G(u) \doteq A\ \mathrm{erfc}(u + a) - B\ \mathrm{erfc}(u + b)$$

for real constants A, B, a, and b, where $A > B > 0$. Let G' be its derivative. Prove the following:

i. If $a \leq b$, then $G(u) > 0$ for all $u \in \mathbb{R}$.

ii. If $a > b$, then there exist unique values u_0 and u_1 such that $G(u_0) = 0$ and $G'(u_1) = 0$; furthermore, $u_0 \neq u_1$. [*Hint:* Sketch G, taking into consideration its limit as $u \to \pm\infty$, as well as the sign of $G'(u)$ at all values of u.]

c. Consider the special case in which there exist values s_- and s_+ such that, for all i,

$$\psi_i = \begin{cases} s_- & \text{if } z_i = -1 \\ s_+ & \text{if } z_i = +1. \end{cases}$$

Find a number τ_0, as a function of s_-, s_+, and β, such that the following hold for all $\tau' < 1$:

i. If $\tau' \geq \tau_0$, then for all α, $\Pi(\alpha,\ \tau') \neq \delta m$.

ii. If $\tau' < \tau_0$, then there exists a unique α such that $\Pi(\alpha,\ \tau') = \delta m$. However, the pair $\langle \alpha, \tau' \rangle$ does not satisfy equation (14.19).

d. Let $z_1, \ldots, z_m$ and δ be given as above. Find values for $\psi_1, \ldots, \psi_m$, $c > 0$, and $\tau \in [0, 1-c)$ for which the *only* BrownBoost solutions are when $\tau' = 1$.

14.6 Prove that theorem 14.1 is false when $c = 0$ in the following general sense: Let $z_1, \ldots, z_m$ and $\delta \in \left(0,\ \frac{1}{2}\right)$ be as in exercise 14.5, and let $c = 0$. Find values for $\psi_1, \ldots, \psi_m$ and $\tau \in [0, 1)$ for which *no* BrownBoost solution exists.

14.7 In section 14.2.2, we saw that if BrownBoost terminates at some time $\tau \geq 1 - c$, then the training error of H is bounded by equation (14.26). In this exercise, we will prove a better bound when a randomized version of H is used instead.

Suppose BrownBoost halts, and in addition to the notation of algorithm 14.1, let us write τ and ψ_i for the values of these variables upon termination. Given x, we redefine H to make a random prediction that is $+1$ with probability

$$\frac{\Phi(-F(x),\ \tau)}{\Phi(F(x),\ \tau)+\Phi(-F(x),\ \tau)},$$

and -1 otherwise, where $F(x) \doteq \sum_{r=1}^{R} \alpha_r h_r(x)$.

a. Give an exact expression for the expected training error of H in terms of the potentials of the chips, where expectation is with respect to the randomized predictions of H.

b. For all $\psi \in \mathbb{R}$, show that $\Phi(\psi,\ \tau)+\Phi(-\psi,\ \tau) \geq 2\Phi(0,\ \tau)$.

c. Show that the expected training error of H is at most

$$\frac{\epsilon}{\operatorname{erfc}\left(\sqrt{\beta c/2}\right)},$$

which approaches ϵ as $c \to 0$.

14.8 Using the notation and assumptions of theorem 14.2 and its proof, this exercise explores what happens if BrownBoost is *not* run lingeringly. We suppose that this happens for the first time on round r, meaning that the preceding $r-1$ rounds were run lingeringly. In particular, this implies that equations (14.29) and (14.30) are still valid. We assume further that $\psi_i^* > 0$ for all i.

a. For $a \in (0, c)$ and $\beta > 0$, let

$$\mathcal{L}_a^\beta \doteq \left\{\langle \alpha, \tau' \rangle : \Pi^\beta(\alpha,\ \tau') = \Pi^\beta(0,\ \tau^\beta),\ 1-c \leq \tau' \leq 1-a\right\}.$$

Show that for all β sufficiently large, there exists $a \in (0, c)$ for which $\mathcal{L}_a^\beta$ is nonempty and compact. [*Hint:* To show non-emptiness, first argue that $\Pi^\beta(0,\ 1) < m\epsilon$, but $\Pi^\beta(0,\ 1-c) > m\epsilon$.]

b. Using part (a), show that for β sufficiently large, there exists a pair $\langle \alpha^\beta, \tau'^\beta \rangle$ which satisfies both equations (14.33) and (14.34), and where $1-c < \tau'^\beta < 1$.

c. We assume henceforth that $\langle \alpha^\beta, \tau'^\beta \rangle$ are chosen as in part (b). Let q be any constant in $(0, 1)$. For β sufficiently large, show that $\tau'^\beta > 1-\beta^{q-2}$. [*Hint:* Adapt the proof of lemma 14.4.]

d. Let

$$\mu_+ \doteq \min_{i:z_i=+1} \psi_i^*, \quad \mu_- \doteq \min_{i:z_i=-1} \psi_i^*,$$

and let $\tilde{\alpha} \doteq (\mu_- - \mu_+)/2$. Show that $\alpha^\beta \to \tilde{\alpha}$ as $\beta \to \infty$. Is this limit $\tilde{\alpha}$ necessarily equal to α^*? Justify your answer. [*Hint:* For all $\delta > 0$ and for β large, show that if $|\alpha^\beta - \tilde{\alpha}| \geq \delta$, then equation (14.34) cannot be satisfied.]