Appendix

Joseph K. Bradley
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
jkbradle@cs.cmu.edu

Robert E. Schapire
Department of Computer Science
Princeton University
Princeton, NJ 08540
schapire@cs.princeton.edu

1 Proof of Theorem 1

Let $\pi(F+\alpha h) = \mathrm{E}[\ln(1+e^{-y(F(x)+\alpha h(x))})]$. Given the previous estimate $F(x)$, we first fix α and choose $h(x)$ to minimize a second-order expansion of $\pi(F+\alpha h)$ around $h(x)=0$.

$$\begin{aligned}\pi(F+\alpha h) &= \mathrm{E}\left[\ln(1+e^{-yF(x)}) - \frac{y\alpha h(x)}{1+e^{yF(x)}} + \frac{1}{2}\frac{y^2\alpha^2 h(x)^2 e^{yF(x)}}{(1+e^{yF(x)})^2}\right] \\ &= \mathrm{E}\left[\ln(1+e^{-yF(x)}) - \frac{y\alpha h(x)}{1+e^{yF(x)}} + \frac{1}{2}\frac{\alpha^2 e^{yF(x)}}{(1+e^{yF(x)})^2}\right]\end{aligned}$$

For $\alpha > 0$, minimizing this approximation of $\pi(F+\alpha h)$ with respect to $h(x)$ is equivalent to maximizing the weighted expectation $\mathrm{E}_q[yh(x)] \equiv \mathrm{E}[q(x,y)yh(x)]$ where $q(x,y) = \frac{1}{1+e^{yF(x)}}$. This criterion is optimized for $f(x) = \mathrm{sign}(\mathrm{E}_q[y|x])$.

Now, given $h(x)$, FilterBoost chooses α to minimize the upper bound

$$\pi(F+\alpha h) \le \mathrm{E}[e^{-y(F(x)+\alpha h(x))}].$$

This is the same optimization objective used by AdaBoost and is minimized when $\alpha = \frac{1}{2}\log(\frac{1/2+\gamma}{1/2-\gamma})$ where γ is the edge of $h(x)$; this is exactly the α used by FilterBoost. ■

2 Proof of Lemma 1

$$\pi_t - \pi_{t+1} = \sum_{(x,y)} D(x,y)\ln\left(\frac{1-q_{t+1}(x,y)}{1-q_t(x,y)}\right) \tag{1}$$

Since $q_t(x,y) = \frac{1}{1+e^{yF_t(x)}}$, $F_t(x) = \sum_{t'=1}^{t-1}\alpha_{t'}h_{t'}(x)$,

$$e^{yF_t(x)} = \frac{1}{q_t(x,y)} - 1 \text{ and} \tag{2}$$

$$q_{t+1}(x,y) = \frac{1}{1+e^{yF_t(x)+\alpha_t y h_t(x)}} \tag{3}$$

Defining $v_t(x,y) = \alpha_t y h_t(x)$, combining (2) and (3) gives

$$q_{t+1}(x,y) = \frac{1}{1+(\frac{1}{q_t(x,y)}-1)e^{v_t(x,y)}} = \frac{q_t(x,y)}{q_t(x,y)+(1-q_t(x,y))e^{v_t(x,y)}} \tag{4}$$

Substituting (4) into (1), and using $\ln(1+z) \le z$, gives

$$\begin{aligned}\pi_t - \pi_{t+1} &= -\textstyle\sum_{(x,y)} D(x,y)\ln(q_t(x,y)e^{-v_t(x,y)} + 1 - q_t(x,y)) \\ &\ge -\textstyle\sum_{(x,y)} D(x,y)(-q_t(x,y) + q_t(x,y)e^{-v_t(x,y)}) \\ &= \textstyle\sum_{(x,y)} D(x,y)q_t(x,y) - \sum_{(x,y)} D(x,y)q_t(x,y)e^{-v_t(x,y)}\end{aligned}$$

Let $D_t(x,y) = \frac{D(x,y)q_t(x,y)}{p_t}$. Then we can write

$$\pi_t - \pi_{t+1} \geq p_t - p_t \sum_{(x,y)} D_t(x,y) e^{-\alpha_t y h_t(x)} \tag{5}$$

Using $\alpha_t = \frac{1}{2}\ln(\frac{1/2+\gamma_t}{1/2-\gamma_t})$ and $\epsilon_t \equiv \Pr_{D_t}[\text{sign}(h_t(x)) \neq y]$ lets us write

$$\sum_{(x,y)} D_t(x,y) e^{-\alpha_t y h_t(x)} = e^{-\alpha_t}(1-\epsilon_t) + e^{\alpha_t}\epsilon_t = 2\sqrt{\frac{1}{4} - \gamma_t^2}$$

Substituting this factor into (5) completes the proof. ■

3 Proof of Theorem 4

Suppose $p_t > \varepsilon/2$. Then the probability that the filter rejects n sequential examples is $(1-p_t)^n < (1-\varepsilon/2)^n$. So, if $(1-\varepsilon/2)^n \leq \delta_t'$, then $p_t \leq \varepsilon/2$ with probability at least $1-\delta_t'$. From Theorem 2, we know $p_t \leq \varepsilon/2$ implies $err_t \leq \varepsilon$. The condition $(1-\varepsilon/2)^n \leq \delta_t'$ gives our bound on n to ensure $err_t \leq \varepsilon$ with high probability. ■

4 Proof of Lemma 2

The proof is identical to Lemma 1 up to (5). Now, though, $\alpha_t = \frac{1}{2}\ln(\frac{1/2+\hat{\gamma}_t'}{1/2-\hat{\gamma}_t'})$. Using $\Pr[|\hat{\gamma}_t - \gamma_t| \leq \tau\gamma_t] > 1-\delta_t$ and $\hat{\gamma}_t' = \frac{\hat{\gamma}_t}{1+\tau}$, we know $\gamma_t \geq \frac{\hat{\gamma}_t}{1+\tau}$ with probability at least $1-\delta_t$, which in turn implies $\hat{\gamma}_t' \leq \gamma_t$. So we may rewrite and bound the sum in (5) as:

$$\begin{aligned}
&\sum_{(x,y)} D_t(x,y) e^{-\alpha_t y h_t(x)} = e^{-\alpha_t}(1-\epsilon_t) + e^{\alpha_t}\epsilon_t \\
= &\left(\frac{\frac{1}{2}-\hat{\gamma}_t'}{\frac{1}{2}+\hat{\gamma}_t'}\right)^{1/2}\left(\frac{1}{2}+\gamma_t\right) + \left(\frac{\frac{1}{2}+\hat{\gamma}_t'}{\frac{1}{2}-\hat{\gamma}_t'}\right)^{1/2}\left(\frac{1}{2}-\gamma_t\right) \\
\leq &\left(\frac{\frac{1}{2}-\hat{\gamma}_t'}{\frac{1}{2}+\hat{\gamma}_t'}\right)^{1/2}\left(\frac{1}{2}+\hat{\gamma}_t'\right) + \left(\frac{\frac{1}{2}+\hat{\gamma}_t'}{\frac{1}{2}-\hat{\gamma}_t'}\right)^{1/2}\left(\frac{1}{2}-\hat{\gamma}_t'\right) \\
= &\ 2\sqrt{1/4 - \hat{\gamma}_t'^2}
\end{aligned}$$

Substituting in $\hat{\gamma}_t' = \frac{\hat{\gamma}_t}{1+\tau}$ and using $\hat{\gamma}_t \geq \gamma_t(1-\tau)$ gives

$$\begin{aligned}
\textstyle\sum_{(x,y)} D_t(x,y) e^{-\alpha_t y h_t(x)} &\leq 2\sqrt{1/4 - (\tfrac{\hat{\gamma}_t}{1+\tau})^2} \\
&\leq 2\sqrt{1/4 - \gamma_t^2(\tfrac{1-\tau}{1+\tau})^2}
\end{aligned}$$

Substituting into (5) gives the required bound. ■

5 Datasets

Majority is generated by a majority vote rule among 40 of 100 binary attributes, with labels corrupted with 10% probability. Twonorm is a noisy synthetic dataset with 20 real-valued attributes from Breiman (1998). Adult is from the UCI Machine Learning Repository (Newman et al., 1998, donated by Ron Kohavi). Adult consists of 14-attribute census data, with labels indicating income level, and eliminating examples with missing attribute values left 45222 examples. Covertype (copyrighted by Jock A. Blackard and Colorado State U.) is also from the UCI Machine Learning Repository. It contains 54-attribute forestry data, where examples are locations and labels indicate the type of tree cover. The original dataset has 7 classes, but we combined the 6 smallest to make the dataset binary, leaving the largest (49% of the examples) alone.