Optimal and Adaptive Algorithms for Online Boosting
Supplementary Material

A. Proof of Lemma 1

Proof. Fix a weak learner, say WL^i. Let

$$U = \{t : (\mathbf{x}_t, y_t) \text{ passed to } \mathrm{WL}^i\}.$$

Since inequality (1) holds even for *adaptive* adversaries, with high probability we have

$$\sum_{t=1}^{T} \mathbf{1}\{\mathrm{WL}^i(\mathbf{x}_t) \neq y_t\}\mathbf{1}\{t \in U\} \leq (\tfrac{1}{2} - \gamma)|U| + S. \quad (1)$$

Now fix the internal randomness of WL^i. Note that $\mathbb{E}_t[\mathbf{1}\{t \in U\}] = p_t^i = \frac{w_t^i}{\|\mathbf{w}^i\|_\infty}$, where $\mathbb{E}_t[\cdot]$ is the expectation conditioned on all the randomness of the booster until (and not including) round t. Define $\sigma = \sum_{t=1}^T p_t^i$.

We now show using martingale concentration bounds that with high probability,

$$\begin{aligned}&\sum_{t=1}^{T} \mathbf{1}\{\mathrm{WL}^i(\mathbf{x}_t) \neq y_t\}p_t^i \\ &\leq \sum_{t=1}^{T} \mathbf{1}\{\mathrm{WL}^i(\mathbf{x}_t) \neq y_t\}\mathbf{1}\{t \in U\} + \tilde{O}\left(\sqrt{\sigma}\right) \quad (2)\end{aligned}$$

and

$$|U| \leq \sigma + \tilde{O}\left(\sqrt{\sigma}\right). \quad (3)$$

Here, the $\tilde{O}(\cdot)$ notation suppresses dependence on $\log\log(T)$.

To prove inequality (2), consider the martingale difference sequence

$$X_t = \mathbf{1}\{\mathrm{WL}^i(\mathbf{x}_t) \neq y_t\}\mathbf{1}\{t \in U\} - \mathbf{1}\{\mathrm{WL}^i(\mathbf{x}_t) \neq y_t\}p_t^i.$$

Note that $|X_t| \leq 1$, and the conditional variance satisfies

$$\mathrm{Var}_t[X_t | X_1, X_2, \ldots, X_{t-1}] \leq p_t^i.$$

Then, by Lemma 2 of Bartlett et al. (2008), for any $\delta < 1/e$ and assuming $T \geq 4$, with probability at least $1 - \log_2(T)\delta$, we have

$$\sum_{t=1}^{T} X_t \leq 2\max\left\{2\sqrt{\sigma}, \sqrt{\ln(\tfrac{1}{\delta})}\right\}\sqrt{\ln(\tfrac{1}{\delta})} = \tilde{O}(\sqrt{\sigma}),$$

by choosing $\delta \ll \frac{1}{\log_2(T)}$. This implies inequality (2). Inequality (3) is proved similarly. Note that these high probability bounds are conditioned on the internal randomness of WL^i. By taking an expectation of this conditional probability over the internal randomness of WL^i, we conclude that inequalities (2) and (3) hold with high probability unconditionally.

Via a union bound, inequalities (1), (2) and (3) all hold simultaneously with high probability, which implies that

$$\sum_{t=1}^{T} \mathbf{1}\{\mathrm{WL}^i(\mathbf{x}_t) \neq y_t\}p_t^i \leq (\tfrac{1}{2} - \gamma)\sigma + S + \tilde{O}\left(\sqrt{\sigma}\right). \quad (4)$$

Using the facts that $p_t^i = \frac{w_t^i}{\|\mathbf{w}^i\|_\infty}$ and $\mathbf{1}\{\mathrm{WL}^i(\mathbf{x}_t) \neq y_t\} = \frac{1 - z_t^i}{2}$ and simplifying, we get

$$\begin{aligned}\mathbf{w}^i \cdot \mathbf{z}^i &\geq 2\gamma\|\mathbf{w}^i\|_1 - 2S\|\mathbf{w}^i\|_\infty - \tilde{O}(\sqrt{\|\mathbf{w}^i\|_1\|\mathbf{w}^i\|_\infty}) \\ &\geq 2\gamma\|\mathbf{w}^i\|_1 - 2S\|\mathbf{w}^i\|_\infty - \gamma\|\mathbf{w}^i\|_1 - \tilde{O}(\tfrac{\|\mathbf{w}^i\|_\infty}{\gamma}) \\ &= \gamma\|\mathbf{w}^i\|_1 - 2S\|\mathbf{w}^i\|_\infty - \tilde{O}(\tfrac{\|\mathbf{w}^i\|_\infty}{\gamma}).\end{aligned}$$

The second inequality above follows from the arithmetic mean-geometric mean inequality. This gives us the desired bound. The high probability bound for all weak learners follows by taking a union bound. □

B. Proof of Lemma 4

Proof. Let $X \sim B(m, p)$ be a binomial random variable where $m = N - i$ and $p = 1/2 + \gamma/2$. Also let $q = 1 - p$ and F_X be the CDF of X. By the definition of w_t^i, we have $w_t^i \leq \frac{1}{2}\max_k \Pr\{X = k\}$. We will approximate X by a Gaussian random variable $G \sim N(mp, mpq)$ with density function f and CDF F_G. Note that

$$\begin{aligned}&|\Pr\{X = k\} - \int_{k-1}^{k} f(G)dG| \\ &= |\,(F_X(k) - F_X(k-1)) - (F_G(k) - F_G(k-1))\,| \\ &\leq |F_X(k) - F_G(k)| + |F_X(k-1) - F_G(k-1)|.\end{aligned}$$

So by applying the Berry-Esseen theorem to the above two CDF differences between X and G, we arrive at

$$\left|\Pr\{X = k\} - \int_{k-1}^{k} f(G)dG\right| \leq \frac{2C(p^2 + q^2)}{\sqrt{mpq}},$$

where C is the universal constant stated in the Berry-Esseen theorem. It remains to point out that

$$\begin{aligned}\Pr\{X=k\} &\le \int_{k-1}^{k} f(G)dG + \frac{2C(p^2+q^2)}{\sqrt{mpq}} \\ &\le \max_{G\in R} f(G) + \frac{2C(p^2+q^2)}{\sqrt{mpq}} \\ &= \frac{1}{\sqrt{2\pi mpq}} + \frac{2C(p^2+q^2)}{\sqrt{mpq}} = O\left(\frac{1}{\sqrt{m}}\right),\end{aligned}$$

since $pq = 1/4 - \gamma^2/4 \ge 3/16$. □

C. Proof of Theorem 3

Proof. The proof of both lower bounds use a similar construction. In either case, all examples' labels are generated uniformly at random from $\{-1, 1\}$, and in time period t, each weak learner outputs the correct label y_t independently of all other weak learners and other examples with a certain probability p_t to be specified later. Thus, for any T, by the Azuma-Hoeffding inequality, with probability at least $1-\delta$, the predictions $\hat{y}_t$ made by the weak learner satisfy

$$\begin{aligned}\sum_{t=1}^{T} \mathbf{1}\{y_t \neq \hat{y}_t\} &\le \sum_{t=1}^{T}(1-p_t) + \sqrt{2T\ln(\tfrac{1}{\delta})} \\ &\le \sum_{t=1}^{T}(1-p_t) + \gamma T + \frac{\ln(\frac{1}{\delta})}{2\gamma} \qquad (5)\end{aligned}$$

where the last inequality follows by the arithmetic mean-geometric mean inequality. We will now carefully choose p_t so that inequality (5) implies inequality (1).

For the lower bound on the number of weak learners, we set $p_t = \frac{1}{2} + 2\gamma$, so that inequality (5) implies that with probability at least $1-\delta$, the predictions $\hat{y}_t$ made by the weak learner satisfy

$$\sum_{t=1}^{T} \mathbf{1}\{y_t \neq \hat{y}_t\} \le (\tfrac{1}{2} - \gamma)T + \frac{\ln(\frac{1}{\delta})}{2\gamma} \le (\tfrac{1}{2} - \gamma)T + S.$$

Thus, the weak online learner has edge γ with excess loss S. In this case, the Bayes optimal output of a booster using N weak learners is to simply take a majority vote of all the weak learners (see for instance Schapire & Freund, 2012, Chap. 13.2.6), and the probability that the majority vote is incorrect is $\Theta(\exp(-8N\gamma^2))$. Setting this error to ϵ and solving for N gives the desired lower bound.

Now we turn to the lower bound on the sample complexity. We divide the whole process into two phases: for $t \le T_0 = \frac{S}{4\gamma}$, we set $p_t = \frac{1}{2}$, and for $t > T_0$, we set $p_t = \frac{1}{2} + 2\gamma$. Now, if $T \le T_0$, inequality (5) implies that with probability at least $1-\delta$, the predictions $\hat{y}_t$ made by the weak learner satisfy

$$\sum_{t=1}^{T} \mathbf{1}\{y_t \neq \hat{y}_t\} \le (\tfrac{1}{2}+\gamma)T + \frac{\ln(\frac{1}{\delta})}{2\gamma} \le (\tfrac{1}{2}-\gamma)T + S \quad (6)$$

using the fact that $T \le T_0 = \frac{S}{4\gamma}$ and $S \ge \frac{\ln(\frac{1}{\delta})}{\gamma}$. Next, if $T > T_0$, let $T' = T - T_0$, and again inequality (5) implies that with probability at least $1-\delta$, the predictions $\hat{y}_t$ made by the weak learner satisfy

$$\begin{aligned}\sum_{t=1}^{T} \mathbf{1}\{y_t \neq \hat{y}_t\} &\le \tfrac{1}{2}T_0 + (\tfrac{1}{2} - 2\gamma)T' + \gamma T + \frac{\ln(\frac{1}{\delta})}{2\gamma} \\ &= (\tfrac{1}{2} - \gamma)T + 2\gamma T_0 + \frac{\ln(\frac{1}{\delta})}{2\gamma} \le (\tfrac{1}{2} - \gamma)T + S, \qquad (7)\end{aligned}$$

since $S \ge \frac{\ln(\frac{1}{\delta})}{\gamma}$. Inequalities (6) and (7) imply that the weak online learner has edge γ with excess loss S.

However, in the first phase (i.e. $t \le T_0$), since the predictions of the weak learners are uncorrelated with the true labels, it is clear that no matter what the booster does, it makes a mistake with probability $\frac{1}{2}$. Thus, it will make $\Omega(T_0)$ mistakes with high probability in the first phase, and thus to achieve ϵ error rate, it needs at least $\Omega(T_0/\epsilon) = \Omega(\frac{S}{\epsilon\gamma})$ examples. □

D. Proof of Lemma 5

Proof. It suffice to prove the bound for $\sigma \ge \frac{1}{2}$; the bound for $\sigma < \frac{1}{2}$ follows by symmetry simply changing the sign of α. For $\sigma \in [0.5, 0.95]$, setting $\alpha = \frac{1}{2}\ln(\frac{\sigma}{1-\sigma}) \in [-2, 2]$ gives

$$\sigma e^{-\alpha} + (1-\sigma)e^{\alpha} = \sqrt{4\sigma(1-\sigma)} \le 1 - \tfrac{1}{2}(2\sigma - 1)^2,$$

since $\sqrt{1-x} \le 1 - \frac{1}{2}x$ for $x \in [0, 1]$. For $\sigma \in (0.95, 1]$, setting $\alpha = \frac{1}{2}\ln(\frac{0.95}{0.05}) \in [-2, 2]$ we have

$$\begin{aligned}&\sigma e^{-\alpha} + (1-\sigma)e^{\alpha} \le 0.95e^{-\alpha} + 0.05e^{\alpha} = \sqrt{0.19} \\ &\le \tfrac{1}{2} \le 1 - \tfrac{1}{2}(2\sigma - 1)^2.\end{aligned}$$

□

E. Description of Data Sets

The datasets come from the UCI repository, KDD Cup challenges, and the HCRC Map Task Corpus. Below, d is the number of unique features in the dataset, and s is the average number of features per example.

Dataset	instances	s	d
20news	18,845	93.9	101,631
a9a	48,841	13.9	123
activity	165,632	18.5	20
adult	48,842	12.0	105
bio	145,750	73.4	74
census	299,284	32.0	401
covtype	581,011	11.9	54
letter	20,000	15.6	16
maptaskcoref	158,546	40.4	5,944
nomao	34,465	82.3	174
poker	946,799	10.0	10
rcv1	781,265	75.7	43,001
vehv2binary	299,254	48.6	105

References

Bartlett, Peter L., Dani, Varsha, Hayes, Thomas, Kakade, Sham, Rakhlin, Alexander, and Tewari, Ambuj. High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pp. 335–342, 2008.

Schapire, Robert E. and Freund, Yoav. *Boosting: Foundations and Algorithms*. MIT Press, 2012.