CS 4540/CS 8803: Algorithmic Theory of Intelligence Fall 2023

Lecture 6: Boosting, Concentration Inequalities

Instructor: Santosh Vempala *Lecture date: 9/25, 9/27*

1 Boosting

Definition 1 (Weak Learner(H, γ)). *A weak learner for a hypothesis class H, with parameter $\gamma > 0$, is a procedure that achieves the following. On any distribution D, it finds a hypothesis h that is correct on $1/2 + \gamma$ of D.*

Our goal is to find a strong learner that outputs a hypothesis that classifies $1 - \epsilon$ of D correctly. The following procedure converts a series of weak learners to a strong learner:

Algorithm 1 Boosting

Start with $w_i = 1$, for $i = 1 \ldots m$.
Draw m samples from D.
for $t = 1 \ldots T$ **do**
  Run the weak learner on the discrete distribution defined by the m samples weighted by w_i.
  Obtain hypothesis h_t, which achieves accuracy $1/2 + \gamma$.
  For all incorrect i, increase $w_i \to w_i \left(\frac{1/2+\gamma}{1/2-\gamma}\right)$
end for
Finally, output the hypothesis which takes the majority of all T, $\text{MAJ}(h_1, \ldots, h_T)$.

Theorem 2. *Algorithm 1 (ϵ, δ) PAC-learns H provided that $T \geq \frac{\ln m}{2\gamma^2}$ and $m \geq \frac{c}{\epsilon}\left(\frac{\ln m}{\gamma^2} \ln H(2m) + \ln \frac{1}{\delta}\right)$.*

Proof. Suppose that $MAJ(h_1, \ldots, h_T)$ makes M_1 mistakes. At this point, for each i with a mistake, $w_i \geq \left(\frac{1/2+\gamma}{1/2-\gamma}\right)^{T/2}$.

Let W_t be the total weight at time t. So, $W_T \geq M_1 \left(\frac{1/2+\gamma}{1/2-\gamma}\right)^{T/2}$.

From t to $t+1$, each mistake is increased by $\frac{1/2+\gamma}{1/2-\gamma}$. The area of the distribution that the mistakes are made on is bounded by $1/2 - \gamma$, per the guarantee of the weak learner. So,

$$W_{t+1} \leq \left(\frac{1/2+\gamma}{1/2-\gamma}\right)(1/2-\gamma)W_t + (1/2+\gamma)W_t = (1+2\gamma)W_t$$

This implies that $W_t \leq m(1+2\gamma)^T$, since the initial weight is m.

Combining the upper and lower bounds,

$$M_1 \left(\frac{1+2\gamma}{1-2\gamma}\right)^{T/2} \leq m(1+2\gamma)^T$$

$$M_1 \leq m(1+2\gamma)^{T/2}(1-2\gamma)^{T/2} = m(1-4\gamma^2)^{T/2}$$

When $T > \frac{\ln m}{2\gamma^2}$, this guarantees that $M_1 < 1$ (meaning that MAJ makes no mistakes).

What about m? A natural candidate would be the sample complexity of PAC-learning H. However, this is not quite sufficient. The hypothesis that this algorithm outputs, $MAJ(h_1, \ldots, h_T)$, is not an element of H, so the sample complexity theorem does not apply.

We can bybass this with the following lemma:

Lemma 3. *Let H be a hypothesis class. Let $H_{MAJ(T)}$ be the class of hypotheses which take the majority of T hypothesis of H. Recall the definition of $H(m)$ from Lecture 3. Then:*

$$H_{MAJ(T)}(m) \leq H(m)^T$$

Proof. Suppose we have any m samples. There are at most $H(m)$ distinct ways to label the samples using one hypothesis of H. Therefore, there are at most $H(m)^T$ ways to pick T distinct labellings of m samples.

For all $\hat{h} \in H_{MAJ(T)}(m)$, $\hat{h}$ is a function of T labellings from H. So, the number of ways to label m samples with $\hat{h} \in H_{MAJ(T)}(m)$ is at most $H(m)^T$. □

Then, we can use the sample complexity theorem (Theorem 6 of Lecture 3, PAC Learning) to bound the number of samples. Theorem 6 gives:

$$m \geq \frac{c}{\epsilon}\left(\ln H_{MAJ(T)}(2m) + \ln \frac{1}{\delta}\right)$$

Substituting the upper bound from the lemma:

$$m \geq \frac{c}{\epsilon}\left(T \ln H(2m) + \ln \frac{1}{\delta}\right)$$

Finally, substituting the value of T:

$$m \geq \frac{c}{\epsilon}\left(\frac{\ln m}{\gamma^2} \ln H(2m) + \ln \frac{1}{\delta}\right)$$

□

Also note that the above lemma can be useful in general for bounding the sample complexity of hypothesis classes that consist of functions of hypotheses from another class.

2 Concentration Inequalities Continued

An important motivating question in concentration inequalities is the following:

Let $X = \sum_{i=1}^{n} X_i$ be a sum of independent random coins, $X_i \in \{0,1\}$, with $Pr(X_i) = \frac{1}{2}$. We want to bound the probability that X deviates from its expectation, $\mathbb{E}X = n/2$.

The fundamental theorem we will use to show this is Markov's inequality:

Theorem 4 (Markov's Inequality)**.** *Let $X \geq 0$, $\mathbb{E}X < \infty$. Then, for any $t > 0$.*

$$Pr(X \geq t\mathbb{E}X) \leq \frac{1}{t}$$

Proof. Remember the definition of expected value; if $Pr(X \geq T) > \frac{1}{t}$, then $\mathbb{E}X > \frac{T}{t}$. Substituting $T = t\mathbb{E}X$ gives a contradiction. □

Another useful inequality uses the variance to give a bound that is often tighter.

Theorem 5 (Chebyshev's Inequality)**.** *Let X be a random variable with $\mathbb{E}(X)$, $Var(X) < \infty$. For any $t > 0$:*

$$Pr(|X - \mathbb{E}X| \geq t\sqrt{Var(X)}) \leq \frac{1}{t^2}$$

Proof. Let $Y = (X - \mathbb{E}X)^2$, and apply Markov's Inequality □

However, for the $X = \sum_{i=1}^{n} X_i$, the sum of random coins, we can get a much tighter bound.

Theorem 6 (Chernoff Bound). *Let $X = \sum_{i=1}^n X_i$ be a sum of independent random indicators ($X_i \in \{0,1\}$) with mean μ. Then, for any $\delta \geq 0$:*

$$\mathbb{P}(X > (1+\delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{2+\delta}\right)$$

For the case of random coins with equal probabilities $Pr(X_i) = \frac{1}{2}$, it is possible to analyze how it decays. Note that $Pr(X = l) = \binom{n}{l}\frac{1}{2^l}$. What is the asymptotic value of $Pr(X = n/2)$? What about $Pr(X = n/2 + k)$?

Before proving the theorem, we first show that the probability that $X = n/2 + k$ is exponentially decreasing with respect to k in the following lemma.

Lemma 7.

$$\mathbb{P}(X = n/2 + k) \leq e^{-k^2/n}$$

Proof. Since X_i is 1 with probability 1/2 and 0 with probability 1/2, $\mathrm{E}[X] := \mathrm{E}[\sum_{i=1}^n X_i] = n/2$. For $0 \leq k \leq n/2$, we can compute the probability that $X = n/2 + k$ as follows.

$$\begin{aligned}
\mathbb{P}(X = \frac{n}{2} + k) =& \frac{\binom{n}{n/2+k}}{\binom{n}{n/2}} = \frac{(n/2)!(n/2)!}{(n/2+k)!(n/2-k)!} \\
=& \frac{(n/2)(n/2-1)\cdots(n/2-k+1)}{(n/2+k)(n/2+k-1)\cdots(n/2+1)} \\
=& (1 - \frac{k}{n/2+k})(1 - \frac{k}{n/2+k-1})\cdots(1 - \frac{k}{n/2+1}) \\
\leq& \exp\left(-k(\frac{1}{n/2+k} + \frac{1}{n/2+k-1} + \cdots + \frac{1}{n/2+1})\right) \\
\leq& \exp(-k\frac{k}{n/2+k}) \\
=& \exp(-\frac{k^2}{n/2+k}) \\
\leq& e^{-k^2/n}
\end{aligned}$$

The first inequality is implied by $1 + x \leq e^x$. $\exp(x) := e^x$ is the exponential function. The last inequality holds because $k \leq n/2$. $\square$

Lemma 7 shows that $\mathbb{P}(X = n/2 + t\sqrt{n}) \leq e^{-t^2}$. This means that the probability drops super fast (in sub-Gaussian decay rate). To bound the tail rate $\mathbb{P}(X \geq t)$, we can use Chernoff Bound. We state the more general case $0 \leq X_i \leq 1, \mathrm{E}[X]_i = p_i, X = \sum_{i=1}^n X_i$ in the following theorem.

Theorem 8 (Chernoff Bound). *Let X_i be independent random variables satisfying $0 \leq X_i \leq 1, \mathrm{E}[X]_i = p_i$. Let $X = \sum_{i=1}^n X_i$ be the sum, with expectation $\mu := \mathrm{E}[X] = \sum_{i=1}^n p_i$. Then for $\delta > 0$*

$$\mathbb{P}(X > (1+\delta)\,\mathrm{E}[X] \leq e^{-\frac{\delta^2}{2+\delta}\mathrm{E}[X]}$$
$$\mathbb{P}(X < (1-\delta)\,\mathrm{E}[X]) \leq e^{-\frac{\delta^2}{2}\mathrm{E}[X]}$$

Proof. We prove the first inequality here. For a fixed $t > 0$, we have

$$\begin{aligned}
\mathbb{P}(X > (1+\delta)\,\mathrm{E}[X]) =& \mathbb{P}(tX > t(1+\delta)\,\mathrm{E}[X]) \\
=& \mathbb{P}(e^{tX} > e^{t(1+\delta)\,\mathrm{E}[X]}) \\
\leq& \frac{\mathrm{E}\left[e^{tX}\right]}{e^{t(1+\delta)\,\mathrm{E}[X]}} \qquad \triangleright \text{ Markov Inequality}
\end{aligned}$$

We first analyze the numerator.

$$\mathrm{E}\left[e^{tX}\right] = \mathrm{E}\left[e^{t\sum_{i=1}^{n} X_i}\right] = \mathrm{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{tX_i}\right]$$

The term $\mathrm{E}\left[e^{tX_i}\right]$ is maximized when X_i is the Bernoulli distribution that is 1 with probability p_i and 0 with probability $1 - p_i$. This implies

$$\mathrm{E}\left[e^{tX_i}\right] \leq p_i e^t + (1 - p_i) = 1 + p_i(e^t - 1) \leq e^{p_i(e^t - 1)}$$

We use $1 + x \leq e^x$ in the last step. Then we have

$$\mathrm{E}\left[e^{tX}\right] = \prod_{i=1}^{n} \mathrm{E}\left[e^{tX_i}\right] \leq \prod_{i=1}^{n} e^{p_i(e^t-1)} = e^{(e^t-1)\sum_{i=1}^{n} p_i} = e^{(e^t-1)\mu}$$

So we can bound the original probability as

$$\mathbb{P}(X > (1+\delta)\,\mathrm{E}[X]) \leq \frac{\mathrm{E}\left[e^{tX}\right]}{e^{t(1+\delta)\,\mathrm{E}[X]}} \leq \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}} = \left(e^{e^t-1-t(1+\delta)}\right)^{\mu}$$

Taking derivative of $e^t - 1 - t(1+\delta)$ on t gives

$$e^t - (1+\delta) = 0.$$

So it reaches the minimum when $t = \ln(1+\delta)$. The minimum is

$$e^t - 1 - t(1+\delta) = \delta - (1+\delta)\ln(1+\delta) \leq -\frac{\delta^2}{2+\delta}.$$

So we choose $t = \ln(1+\delta)$, and thus have

$$\mathbb{P}(X > (1+\delta)\,\mathrm{E}[X]) \leq e^{(\delta-(1+\delta)\ln(1+\delta))\mu} \leq e^{-\frac{\delta^2}{2+\delta}\mu}$$

□

By Theorem 8 with $p = 1/2, \delta = 2t/\sqrt{n}$, we know

$$\mathbb{P}(X \geq n/2 + t\sqrt{n}) \leq e^{-\frac{4t^2}{2n(1+t/\sqrt{n})}\frac{n}{2}} \simeq e^{-t^2}$$

This implies that the tail of X decays exponentially.