CS 7545: Machine Learning Theory **Fall 2019**

Lecture 5: Martingale and Online Learning Introduction

Lecturer: Jacob Abernethy *Scribe: Wenbo Chen*

5.1 Martingales

In this section we will introduce martingales and prove Azuma's inequality.

Definition 5.1 (Martingale) *A sequence of random variable $Z_0, Z_1, \cdots, Z_n$ is called a* ***martingale*** *sequence if for all $n \in \mathbb{N}$*

- $\mathbb{E}[|Z_n|] < \infty$
- $\mathbb{E}[Z_n|Z_1, \cdots, Z_{n-1}] = Z_{n-1}$

Example 1 (Linear Martingales) Let $X_1, \cdots, X_n$ be a sequence of $i.i.d$ random variable with $\mathbb{E}[X_i] = 0$, for all $i > 0$, then $Z_n = \sum_{i=1}^n X_i$ is a martingale.
Proof:

$$\begin{aligned}
\mathbb{E}(|Z_n|) &= \mathbb{E}(|\sum_{i=1}^n X_i|) < \infty, \forall n \geq 0. && (5.1)\\
\mathbb{E}[Z_n|X_{n-1}, \cdots, X_1] &= \mathbb{E}[\sum_{i=1}^n X_i|X_{n-1}, \cdots, X_1] && (5.2)\\
&= \mathbb{E}[X_n|X_{n-1}, \cdots, X_1] + \mathbb{E}[Z_{n-1}|X_{n-1}, \cdots, X_1] && (5.3)\\
&= \mathbb{E}[X_n] + Z_{n-1} && (5.4)\\
&= Z_{n-1} && (5.5)
\end{aligned}$$

Therefore, according to **Def. 5.1**, Z_n is a martingale. ■

Example 2 (Quadratic Martingales) Let $X_1, \cdots, X_n$ be a sequence of $i.i.d$ random variable with $\mathbb{E}[X_i] = 0$ and $\sigma^2 = \operatorname{var}(X_i) < \infty$. Let $S_n = \sum_{i=1}^n X_n, Z_n = S_n^2 - n\sigma^2$. Then $\{Z_i\}_{i\geq 0}$ is a martingale.
Proof:

$$\begin{aligned}
\mathbb{E}\left[\left|S_n^2 - n\sigma^2\right|\right] &\leq \mathbb{E}\left[S_n^2\right] + n\sigma^2 && (5.6)\\
&= \operatorname{var}(S_n) + \mathbb{E}^2\left[S_n\right] + n\sigma^2 && (5.7)\\
&= n \operatorname{var}(X_1) + (n\mathbb{E}\left[X_i\right])^2 + n\sigma^2 && (5.8)\\
&= 2n\sigma^2 < \infty && (5.9)\\
\mathbb{E}[Z_n|X_1, \cdots, X_{n-1}] &= \mathbb{E}[(S_{n-1} + X_n)^2 - n\sigma^2|X_1, \cdots, X_{n-1}] && (5.10)\\
&= \mathbb{E}[S_{n-1}^2 + X_n^2 + 2S_{n-1}X_n - n\sigma^2|X_1, \cdots, X_{n-1}] && (5.11)\\
&= \mathbb{E}[S_{n-1}^2 - (n-1)\sigma^2|X_1, \cdots, X_{n-1}] && (5.12)\\
&= S_{n-1}^2 - (n-1)\sigma^2 = Z_{n-1} && (5.13)
\end{aligned}$$

Therefore, according to **Def. 5.1**, Z_n is a martingale. ■

Lemma 5.2 (Hoeffding Lemma) *Let X be a bound random variable with $X \in [a, b]$ and $\mathbb{E}[X] = 0$, then X is sub-Gaussian with variance proxy $\frac{(b-a)^2}{4}$ and*

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Lemma 5.3 (Tower Rule) *Let X be a random variable whose expected value $\mathbb{E}(X)$ is defined, and Y be any random variable on the same probability space, then*

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

Theorem 5.4 (Azuma's Inequality) *Let $Z_0, Z_1, \cdots, Z_n$ be a martingale $|Z_i - Z_{i-1}| \leq c_i$, $\forall i \geq 1$. Then*

$$\mathbb{P}(Z_n - Z_0 \geq t) \leq \exp(-\frac{-t^2}{2\sum_{i=1}^n c_i^2})$$

Proof: *Let $s > 0$, we have:*

$$\begin{aligned}
\mathbb{P}(Z_n - Z_0 \geq t) &= \mathbb{P}(\exp(s(Z_n - Z_0)) \geq \exp(st)) && (5.14)\\
&\leq \frac{\mathbb{E}[\exp(s(Z_n - Z_0))]}{\exp(st)} && (5.15)\\
&= \exp(-st)\mathbb{E}[\exp(s(Z_n - Z_{n-1}))\exp(s(Z_{n-1} - Z_0))] && (5.16)\\
&= \exp(-st)\mathbb{E}[\mathbb{E}[\underbrace{\exp(s(Z_n - Z_{n-1}))}_{\text{Hoeffding Lemma}}\ \underbrace{\exp(s(Z_{n-1} - Z_0))}_{\text{constant given } Z_0,\cdots,Z_{n-1}}\ |Z_0, \cdots, Z_{n-1}]] && (5.17)\\
&\leq \exp(-st)\exp(\frac{s^2c_i^2}{2})\mathbb{E}[\exp(s(Z_{n-1} - Z_0))] && (5.18)\\
&\leq \exp(-st)\prod_{i=1}^n \exp(\frac{s^2c_i^2}{2}), && (5.19)
\end{aligned}$$

where we compute the upper bound recursively from (5.18) with Hoeffding Lemma and Tower Rule and then get (5.19). Note that we can use Hoeffding Lemma due to $\mathbb{E}[Z_n|Z_0, \cdots, Z_{n-1}] = Z_{n-1} \Rightarrow \mathbb{E}[Z_n - Z_{n-1}|Z_0, \cdots, Z_{n-1}] = 0$.

Since the bound holds with all $s > 0$, we have:

$$\mathbb{P}(Z_n - Z_0 \geq t) \leq \inf_{s>0} \exp(-st)\prod_{i=1}^n \exp(\frac{s^2c_i^2}{2}) = \exp\left(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\right),$$

where $s = \frac{t}{\sum_{i=1}^n c_i^2}$. ■

5.2 Online Learning Introduction

Example (Weather report) At each round $t \in \mathbb{N}^+$

- N weather experts predict weather $\{x_{i,t}\}_{i\in[N]}$, where for all i, t, $x_{i,t} \in \{0, 1\}$ and 0 indicates no rain and 1 indicates rain, $[N] = \{1, 2, \cdots, N\}$;
- Algorithm predicts $\hat{y}_t \in \{0, 1\}$;
- Nature reveals $y_t \in \{0, 1\}$;
- Assume that there are perfect experts.

Let $\#mistakes$ denotes the number of mistakes the algorithm made before finding the perfect expert, where mistake happens when $\hat{y}_t \neq y_t$. The problem here is, **for a particular algorithm, at most how many mistakes the algorithm needs to make until it finally finds the perfect expert?**

Algorithm 1 Halving Algorithm

1: $\mathcal{C}_1 = [N]$
2: **for** $t = 1, \cdots, T$ **do**
3: observe $x_{i,t}$ $\forall i \in \mathcal{C}_t$
4: $\hat{y}_t = round(\frac{1}{|\mathcal{C}_t|} \sum_{i \in \mathcal{C}_t} x_{i,t})$ ▷ Majority vote
5: $\mathcal{C}_{t+1} = \mathcal{C}_t \backslash \{i : x_{i,t} \neq y_t\}$ ▷ Bad experts elimination
6: **end for**

Theorem 5.5 *Halving algorithm satisfies $\#mistakes \leq \log_2 N$.*
Proof: $|\mathcal{C}_1| = N$. *If $\#mistakes$ increases we have*

$$\frac{|\mathcal{C}_{t+1}|}{|\mathcal{C}_t|} \leq \frac{1}{2} \tag{5.20}$$

$$\Rightarrow \quad 1 \leq |\mathcal{C}_T| \leq |\mathcal{C}_1|(\frac{1}{2})^{\#mistakes} = N(\frac{1}{2})^{\#mistakes} \tag{5.21}$$

$$\Rightarrow \quad 0 \leq \log_2 N - \#mistakes \tag{5.22}$$

$$\Rightarrow \quad \#mistakes \leq \log_2 N \tag{5.23}$$

Therefore, we can see that the number of mistakes is bounded by $\log_2 N$. ■