1 Background on Expectation

Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Expectation: Let $Z : \Omega \rightarrow \mathbb{R}$ be a random variable. $E[Z] = \sum_{z \in \Omega} P(Z = z)z$.

Conditional Expectation 1: $E[Z|Y = y] = \sum_z P(Z = z|Y = y)z$. This expectation is a function of $y$ and hence a number.

Conditional Expectation 2: $E[Z|Y] = \sum_z P(Z = z|Y)z$. This expectation is a function of $Y$ and hence a random variable.

Example 1

$X = \begin{cases} 1 \text{ w.p. } 1/2 \\ 0 \text{ w.p. } 1/2 \end{cases}$

$Y = \begin{cases} 2 \text{ w.p. } 1/2 \\ 0 \text{ w.p. } 1/2 \end{cases}$

Let $Z = X + Y$. Note that, $E[Z] = \frac{1}{2}[0 + 1 + 2 + 3] = 3/2$.

$E[Z|X] = \begin{cases} E[Z|X = 0] \text{ w.p. } P(X = 0) = 1/2 \\ E[Z|X = 1] \text{ w.p. } P(X = 1) = 1/2 \end{cases}$

Note that $E[Z|X] = E[Y] + X = 1 + X$.

Lemma 1 (Law of Total Expectation). $\forall X, Y \ E[X] = E[E[X|Y]]$.

Proof.

\[
E[X] = \sum_x P(X = x)x = \sum_y \left( \sum_x P(X = x, Y = y) \right)x = \sum_y P(Y = y) \sum_x P(X = x|Y = y) x = \sum_y P(Y = y) E[X|Y = y] = E[E[X|Y]]
\]

where the second equality follows from total probability.

Example 2 Let $U_1, U_2, \ldots, U_m$ be random variables. Let $X = f(U_1, U_2, \ldots, U_m)$ and $Z = E[X|U_1, U_2, \ldots, U_k]$. Then, $E[X] = E_{U_1, U_2, \ldots, U_k} E_{U_{k+1}, \ldots, U_m}[X] = E_{U_1, \ldots, U_k} E[X|U_1, U_2, \ldots, U_k] = E[Z]$. The previous expression also follows from the law of total expectation.

2 Background on Martingales

Definition 2. A sequence of random variables $(W_i)_{i=0}^m$ is a martingale w.r.t another sequence of random variables $(U_i)_{i=1}^m$ if

\[
E[|W_i|] < \infty \quad \forall i \quad E[W_{i+1}|U_1, U_2, \ldots, U_i] = W_i
\]
Example 3 Consider a random walk on real line:

\[
\begin{bmatrix}
-1 & 0 \\
0 & 1
\end{bmatrix}
\]

Let \( W_i \) denote the position after \( i \) steps with the initial position \( W_0 = 0 \). Let the random walk be described by:

\[
W_{i+1} = \begin{cases} 
W_i + 1 & \text{w.p. } 1/2 \\
W_i - 1 & \text{w.p. } 1/2 
\end{cases}
\]

(5)

Let,

\[
U_i = \begin{cases} 
1 & \text{w.p. } 1/2 \\
-1 & \text{w.p. } 1/2 
\end{cases}
\]

(6)

Note that, \( W_{i+1} = W_i + U_{i+1}, i = 0, \ldots, m - 1 \). Also, \( (W_i)_{i=0}^m \) is a martingale w.r.t \( (U_i)_{i=1}^m \) since \( \mathbb{E}[|W_i|] < \infty \) and \( \mathbb{E}[W_{i+1} | U_1, U_2, \ldots, U_i] = \mathbb{E}[\sum_{j=1}^{i+1} U_j | U_1, U_2, \ldots, U_i] = \sum_{j=1}^{i+1} U_i + \mathbb{E}[U_{i+1}] = W_i \).

3 Doob Martingale

Definition 3. Let \( (U_i)_{i=1}^m \) be a sequence of random variables and \( f(U_1, U_2, \ldots, U_m) \) be a function such that \( \mathbb{E}[|f(U_1, U_2, \ldots, U_m)|] < \infty \). The doob martingale is defined as \( (W_i)_{i=0}^m \) where, \( W_i = \mathbb{E}[f(U_1, U_2, \ldots, U_m) | U_1, U_2, \ldots, U_i] \) for \( 1 \leq i \leq m \) and \( W_0 = \mathbb{E}[f(U_1, \ldots, U_m)] \).

Note that the randomness is incrementally revealed in \( W_i \) as \( i \) goes from 0 (\( W_0 \) is a scalar) to \( m \) (\( W_m \) is a function of \( U_1, U_2, \ldots, U_m \)).

Theorem 4. A doob martingale is a martingale.

Proof. For a doob martingale, \( \mathbb{E}[|W_i|] = \mathbb{E}[U_1, \ldots, U_i | \mathbb{E}[U_{i+1}, \ldots, U_m | f(U_1, \ldots, U_m)]] \leq \mathbb{E}[|f(U_1, \ldots, U_m)|] < \infty \).

Also, \( \mathbb{E}[W_{i+1} | U_1, \ldots, U_i] = \mathbb{E}[\mathbb{E}[f(U_1, \ldots, U_m) | U_1, \ldots, U_i] | U_1, \ldots, U_i] = \mathbb{E}[U_{i+1}, \ldots, U_m | \mathbb{E}[U_{i+1}, \ldots, U_m | f(U_1, \ldots, U_m)] = \mathbb{E}[U_{i+1}, \ldots, U_m | f(U_1, \ldots, U_m)] = \mathbb{E}[f(U_1, \ldots, U_m) | U_1, \ldots, U_i] = W_i \). \( \square \)

Definition 5. We say that a martingale \( (W_i)_{i=0}^m \) has \( \frac{c}{m} \)-bounded differences (\( \frac{c}{m} \)-Lipschitz) if \( |W_{i+1} - W_i| \leq \frac{c}{m} \).

Fact 6. Hoeffding-Azuma: For Doob Martingales with \( \frac{c}{m} \)-bounded differences, \( \mathbb{P}(W_m - W_0 < \epsilon) \leq e^{-\frac{2m \epsilon^2}{c}} \).

Thus, for all \( \delta > 0 \) w.p. \( 1 - \delta \) over the random draws \( U_1, \ldots, U_m \),

\[
W_m \leq W_0 + c\sqrt{\frac{\log \frac{1}{\delta}}{2m}}
\]

(7)

For general martingales, \( \mathbb{P}(W_m - W_0 > \epsilon) \leq e^{-\frac{m \epsilon^2}{2c^2}} \).

4 Online Learning

Assume the samples, \( S \in \mathcal{D}^m \). Let \( (\mathcal{X}, \mathcal{Y}) \) denote the set of all possible feature vectors and labels respectively. The general form of online learning is as follows:

1. Start with \( h_0 \in \mathcal{H} \).
2. For \( i = 1, 2, \ldots, m \),

\[
\]
(a) Receive \( x_i \in X \).
(b) Predict \( h_{i-1}(x_i) \).
(c) Receive \( y_i \in Y \).
(d) Suffer loss \( l(h_{i-1}; (x_i, y_i)) \).
(e) Update \( h_i \leftarrow A(h_{i-1}; (x_i, y_i)) \) (where \( A \) denotes the online algorithm).

Remarks
Now let \( f(S) = \frac{1}{m} \sum_{i=1}^{m} l(h_{i-1}; D) - \sum_{i=1}^{m} l(h_{i-1}; (x_i, y_i)) \). Define \( \sum_{i=1}^{m} l(h_{i-1}; (x_i, y_i)) \) to be the cumulative loss. We make the following remarks.

1. \( l(h_{i-1}; D) \) is a random variable (since \( h_{i-1} \) is a function of \( (x_j, y_j)_{j=1}^{i-1} \) that are drawn i.i.d from \( D \)).
2. \( l(h_{i-1}; (x_i, y_i)) \) is a random variable (with randomness in \( (x_j, y_j)_{j=1}^{i-1} \)).
3. \( E[l(h_{i-1}; (x_i, y_i))|(x_j, y_j)_{j=1}^{i-1}] = E(x_i, y_i)[l(h_{i-1}; (x_i, y_i))] = l(h_{i-1}; D) \).
4. If \( l \in [0, c] \) then \( W_i = E[f(S)|(x_j, y_j)_{j=1}^{i-1}] \) is a Doob martingale with \( \frac{c}{m} \)-bounded differences.

Theorem 7. \( \forall \delta > 0, \text{w.p.} \geq 1 - \delta \) over the random (i.i.d) sampling of \( S \in D^m \),

\[
f(S) \leq c \sqrt{\frac{\log \left( \frac{1}{\delta} \right)}{2m}} \tag{8}
\]

Proof. Remark 4 states that \( W_i = E[f(S)|(x_j, y_j)_{j=1}^{i-1}] \) is a doob martingale. \( W_0 = E[f(S)] \) and \( W_m = f(S) \). Thus, by Hoeffding-Azuma inequality (Fact 6), it follows that w.p. \( \geq 1 - \delta \),

\[
W_m - W_0 \leq c \sqrt{\frac{\log \left( \frac{1}{\delta} \right)}{2m}} \tag{9}
\]

It also follows from Remark 3 that \( E[f(S)] = 0 \) and hence the theorem follows. \( \square \)

Remark
Let \( \frac{1}{m} [\sum_{i=1}^{m} l(h_{i-1}; D)] \) denote the average risk of \( \{h_0, \ldots, h_{m-1}\} \). Also let \( Q \) be the uniform distribution over \( \{h_0, \ldots, h_{m-1}\} \). Then, the previous theorem implies that w.p. \( \geq 1 - \delta \),

\[
l(Q; D) \leq \frac{1}{m} \sum_{i=1}^{m} l(h_{i-1}; (x_i, y_i)) + c \sqrt{\frac{\log \left( \frac{1}{\delta} \right)}{2m}} \tag{10}
\]