CPSC 550: Machine Learning II	2008/9 Term 2
Lecture 21 — Apr 9, 2009	
Lecturer: Nando de Freitas	*Scribe: Nimalan Mahendran*

21.1 Some Concepts from Previous Lectures

The following inequalities were introduced in previous lectures.

Theorem 21.1 (Doob-Kolmogorov Inequality, [Wil91] and [GS01]). *Let $\{S_n\}$ be a martingale and $\epsilon > 0$.*

$$P\Big(\max_{1\leq i\leq n} |S_i| \geq \epsilon\Big) \leq \frac{1}{\epsilon^2}\mathbb{E}(S_n^2)$$

Theorem 21.2 (Chebyshev Inequality, [Was04]). *Let X be a random variable, $\mu = \mathbb{E}(X)$, $\sigma^2 = \mathbb{V}(X)$ and $\epsilon > 0$.*

$$P\big(|X-\mu| \geq \epsilon\big) \leq \frac{\sigma^2}{\epsilon^2}$$

Example, [Was04] *Let $X_i \sim \text{Bernoulli}(d), 1 \leq i \leq n$, be a sequence of random variables such that each $X_i \in \{0,1\}$ is independent, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be a random variable and $d \in [0,1]$.*

$$\mu = \mathbb{E}(\bar{X}_n) = d$$
$$P\big(|\bar{X}_n - d| > \epsilon\big) \leq \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

Theorem 21.3 (Hoeffding Inequality, [Was04]). *Let $X_{1:n}$ be a sequence of independent random variables, with $\mathbb{E}(X_i) = 0$, $a_i \leq Y_i \leq b_i$ and $\epsilon > 0$. The condition $a_i \leq Y_i \leq b_i$ bounds the support of Y_i to be $[a_i, b_i]$. Then for any $t > 0$,*

$$P\Big(\sum_{i=1}^n X_i \geq \epsilon\Big) \leq e^{-t\epsilon} e^{\frac{t^2}{2}\sum_{i=1}^n (b_i - a_i)^2}$$

A specialization of the Hoeffding inequality to the average of n independent Bernoulli trails gives a tighter bound than the Chebyshev inequality.

Theorem 21.4 (Hoeffding Inequality for Ind. Bernoulli Trials, [Was04]). *Let $X_i \sim \text{Bernoulli}(d), 1 \leq i \leq n$, be a sequence of random variables such that each $X_i \in \{0,1\}$ is independent, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be a random variable and $d \in [0,1]$. Then for any $\epsilon > 0$,*

$$P\Big(|\bar{X}_n - d| > \epsilon\Big) \leq 2e^{-2n\epsilon^2}$$

Example, [Was04] Let $X_i \sim \text{Bernoulli}(d), 1 \leq i \leq n$, be a sequence of random variables such that each $X_i \in \{0,1\}$ is independent, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be a random variable and $d \in [0,1]$. Let $n = 100$ and $\epsilon = 0.2$. Then the bounds given by the Chebyshev inequality and the Hoeffding inequality are as follows

$$\text{Bound(Chebyshev)} = 0.0625$$
$$\text{Bound(Hoeffding)} = 0.00067$$

and it can be seen that the Hoeffding inequality gives a much tighter bound.

21.2 Martingale Convergence Theorem

Theorem 21.5 (Martingale Convergence Theorem). *Let $\{S_m\}$ be a martingale with respect to the filtration $\{X_m\}$. If $\mathbb{E}(S_m^2) < M < \infty$ for some M and all m, then $\exists s$ such that $S_m \longrightarrow^{a.s.} s$.*

21.2.1 The Proof of Theorem 21.5

The proof of Theorem 21.5 is split into four parts.

Part A proves that $\mathbb{E}(S_m^2)$ is non-decreasing.

Part B finds an expression for the set of events where convergence does not occur (the non-convergence set) and an expression for its probability.

Part C proves that $Y_m = S_{m+n} - S_m$ is a martingale with respect to itself.

Part D proves that the probability of the non-convergence set is zero.

Part A

$$\mathbb{E}\big(S_{m+n}^2\big) = \mathbb{E}\big(S_m^2\big) + 2\mathbb{E}\big(S_m(S_{m+n} - S_m)\big) + \mathbb{E}\big((S_{m+n} - S_m)^2\big)$$
$$= \mathbb{E}\big(S_m^2\big) + 2\mathbb{E}\big(S_m\mathbb{E}(S_{m+n} - S_m|x_{1:m})\big) + \mathbb{E}\big((S_{m+n} - S_m)^2\big)$$

where the substitution follows from the Tower Property.

$$= \mathbb{E}\big(S_m^2\big) + 2\mathbb{E}\big(S_m\mathbb{E}(S_{m+n}|x_{1:m})\big) - 2\mathbb{E}\big(S_m\mathbb{E}(S_m|x_{1:m})\big) + \mathbb{E}\big((S_{m+n} - S_m)^2\big)$$
$$= \mathbb{E}\big(S_m^2\big) + 2\mathbb{E}\big(S_m s_m\big) - 2\mathbb{E}\big(S_m s_m\big) + \mathbb{E}\big((S_{m+n} - S_m)^2\big)$$

where the substitution for the second term follows from the property of martingales from the last lecture that $\mathbb{E}\big(S_{m+n}|x_{1:m}\big) = s_m$ and the substitution for the third term comes from the certainty of S_m given $x_{1:m}$.

$$= \mathbb{E}\big(S_m^2\big) + \mathbb{E}\big((S_{m+n} - S_m)^2\big)$$
$$\geq \mathbb{E}\big(S_m^2\big)$$

Therefore, $\mathbb{E}(S_m^2)$ is non-decreasing.

Part B

Definition 1 (Cauchy convergence of sequences). *$\{X_n\}$ is a Cauchy convergent sequence if $\forall\epsilon > 0$, there exists an N such that $|X_m - X_n| < \epsilon, \forall m, n \geq N$.*

[Tre03] gives the definition and some examples for the Cauchy convergence of seqeunces.

Definition 2 (Cauchy convergence of sequences of random variables). *$\{X_n\}$ is a Cauchy convergent sequence of random variables if $\forall\epsilon > 0$ there exists an N such that*

$$P\Big(\big\{\omega \in \Omega : |X_m(\omega) - X_n(\omega)| < \epsilon, \forall m, n \geq N\big\}\Big) = 1$$

Definition 2 is similar to the pointwise convergence of functions, where $\{X_n(\omega)\}$ is the sequence of functions. Pointwise convergence of functions is defined in [Tre03].

Definition 3 (Convergence set). *The convergence set C is the set of events $\omega \in \Omega$ where the sequence with respect to n, $S_n(\omega)$, is Cauchy convergent.*

$$C = \big\{\omega \in \Omega : \{s_n(\omega)\} \text{ is Cauchy convergent}\big\}$$
$$= \big\{\omega \in \Omega : \forall\epsilon > 0, \exists m \text{ s.t. } |s_{m+i}(\omega) - s_{m+j}(\omega)| < \epsilon, \forall i, j \geq 0\big\}$$
$$= \big\{\omega \in \Omega : \forall\epsilon > 0, \exists m \text{ s.t. } |s_{m+i}(\omega) - s_m(\omega)| < \epsilon, \forall i \geq 1\big\}$$

The convergence set C can be rewritten as

$$C = \bigcap_{\epsilon>0} \bigcup_{m} \left\{\omega \in \Omega : |s_{m+i}(\omega) - s_m(\omega)| < \epsilon, \forall i \geq 1\right\}$$

The non-convergence set is $\overline{C}$

$$\begin{aligned}
\overline{C} &= \overline{\bigcap_{\epsilon>0} \bigcup_{m} \left\{\omega \in \Omega : |s_{m+i}(\omega) - s_m(\omega)| < \epsilon, \forall i \geq 1\right\}} \\
&= \bigcup_{\epsilon>0} \bigcap_{m} \left\{\omega \in \Omega : \exists i \geq 1 \text{ s.t. } |s_{m+i}(\omega) - s_m(\omega)| \geq \epsilon\right\}
\end{aligned}$$

The probability of the non-convergence set $\overline{C}$ is

$$P(\overline{C}) = \lim_{\epsilon \to 0} \lim_{m \to \infty} P\Big(\left\{\omega \in \Omega : \exists i \geq 1 \text{ s.t. } |s_{m+i}(\omega) - s_m(\omega)| \geq \epsilon\right\}\Big)$$

Part C

Let $Y_n = S_{m+n} - S_m$.

$$\begin{aligned}
\mathbb{E}(Y_{n+1}|Y_{1:n}) &= \mathbb{E}(\mathbb{E}(Y_{n+1}|s_{1:m+n})|Y_{1:n}) \\
&= \mathbb{E}(\mathbb{E}(S_{m+n+1} - S_m|s_{1:m+n})|Y_{1:n}) \\
&= \mathbb{E}(\mathbb{E}(S_{m+n+1}|s_{1:m+n}) - \mathbb{E}(S_m|s_{1:m+n})|Y_{1:n}) \\
&= \mathbb{E}(s_{m+n} - s_m|Y_{1:n}) \\
&= s_{m+n} - s_m = y_n
\end{aligned}$$

Therefore, Y_n is a martingale with respect to itself.

Part D

The Doob-Kolmogorov inequality can be applied to the martingale Y_n to give

$$P(\max_{1 \leq i \leq n} |S_{m+i} - S_m| \geq \epsilon) \leq \frac{\mathbb{E}(S_{m+i} - S_m)^2}{\epsilon^2}, \text{ as } n \to \infty$$

This expression is equivalent to

$$\begin{aligned}
& P(\exists 1 \leq i \leq n \text{ s.t. } |S_{m+i} - S_m| \geq \epsilon) \\
\leq \quad & \frac{1}{\epsilon^2}\mathbb{E}\big((S_{m+n} - S_m)^2\big) \\
\leq \quad & \frac{1}{\epsilon^2}\Big[M^2 - \mathbb{E}(S_m^2)\Big]
\end{aligned}$$

This inequality holds for all n, including $n = \infty$. It can then be substituted into the expression for the probability of non-convergence, $P(\overline{C})$

$$
\begin{aligned}
P(\overline{C}) &= \lim_{\epsilon \to 0} \lim_{m \to \infty} P\Big(\big\{\omega \in \Omega : \exists i \geq 1 \text{ s.t. } |s_{m+i}(\omega) - s_m(\omega)| \geq \epsilon\big\}\Big) \\
&\leq \lim_{\epsilon \to 0} \lim_{m \to \infty} \frac{M^2 - \mathbb{E}(S_m^2)}{\epsilon^2} \\
&= \lim_{\epsilon \to 0} \frac{0}{\epsilon^2}
\end{aligned}
$$

$\because M$ is an upper bound on $\mathbb{E}(S_m^2)$ for all m and $\mathbb{E}(S_m^2)$ is non-decreasing.

$$
\begin{aligned}
&= 0 \\
P(\overline{C}) &\leq 0 \implies P(\overline{C}) = 0 \implies P(C) = 1
\end{aligned}
$$

Therefore, $S_m \longrightarrow^{a.s.} s$. $\square$

21.3 Hoeffding Inequality for Martingales

Theorem 21.6 (Hoeffding Inequality for Martingales). *Let $\{S_n\}$ be a martingale and $\epsilon > 0$. Then for all n,*

$$P(|S_n - S_{n-1}| \leq B_n) = 1 \implies P(|S_n - S_0| \geq \epsilon) \leq 2e^{-\frac{1}{2}\frac{\epsilon^2}{\sum_i B_i^2}}$$

Bibliography

[GS01] Geoffrey R. Grimmett and David R. Stirzaker. *Probability and Random Processes*. Oxford University Press, August 2001.

[Tre03] William F. Trench. Introduction to real analysis. 2003.

[VC71] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.

[Was04] Larry Wasserman. *All of Statistics : A Concise Course in Statistical Inference (Springer Texts in Statistics)*. Springer, September 2004.

[Wil91] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.