CPSC 550: Machine Learning II **2008/9 Term 2**

Lecture 20 — Apr 2, 2009

Lecturer: Nando de Freitas *Scribe: Bo Chen*

Following previous efforts on proving convergence for stochastic approximation, this lecture completes the final step of the proof by integrating Lyapanov functions, Martingales (section 20.2), conditional expecation and the Doob-Kolmogorov inequality (section 20.3) into a unifying framework (section 20.4).

20.1 Review: Conditional Expecation

The previous lecture introduced the conditional expectation, $\mathbb{E}[X|Z]$, which is an random variable indepedent of X. Let $\mathbb{I}_A(z)$ denote the indicator function over a set A, $\mathbb{I}_A(z) = 1$ iff $z \in A$. Then $\mathbb{E}[X|A]$ is related to $\mathbb{E}[X\mathbb{I}_A]$ in the following equation:

$$\begin{aligned}
\mathbb{E}[X\mathbb{I}_A] &= \iint x\mathbb{I}_A(z)dP_{X,Z}(x,z) \\
&= \iint x\mathbb{I}_A(z)dP_{X|Z}(x|z)dP_Z(z) \\
&= \int xdP_{X|Z}(x|A)P(A) = P(A)\int xdP_{X|Z}(x|A) \\
&= P(A)\mathbb{E}[X|A].
\end{aligned}$$

Example 20.1. *If $A = \bigcup_{i=1}^3 B_i$, then*

$$\mathbb{E}[X\mathbb{I}_A] = \sum_{i=1}^{3}\mathbb{E}[X|B_i]P(B_i) = \sum_{i=1}^{3}\mathbb{E}[X\mathbb{I}_{B_i}].$$

20.2 Martingales

In this section, filtration and adapted processes are introduced, followed by the formal definition of Martingales.

Definition 20.1. *Given a measurable space* $(\Omega, \mathcal{F}, P)$ *and an totally ordered index set* $\mathcal{I}$*, a* **Filtration** *is an increasing family of* **sub-sigma algebras** $\{\mathcal{F}_n\}_{n\in\mathcal{I}}$ *such that:*

$$\mathcal{F}_i \subseteq \mathcal{F}_j \qquad \forall i \leq j \in \mathcal{I}$$

and

$$\mathcal{F}_\infty = \sigma\left(\bigcup_{i\in\mathcal{I}} \mathcal{F}_i\right)$$

Example 20.2. *In a Markov Chain with state variables* $\{S_t\}_{t=1}^T$ *and observation variables* $\{O_t\}_{t=1}^T$*, the filtering operation at time* t *computes the conditional expectation* $\mathbb{E}[S_t|\sigma(O_{1:t})]$*. The sigma field of the observations* $O_{1:t}$ *is a filtration because it grows over time and satisfies*

$$\sigma(O_{1:t_1}) \subseteq \sigma(O_{1:t_2}) \quad \forall t_1 \leq t_2$$

Without loss of generality, $\mathcal{I}$ is henceforth assumed to be the set of non-negative integers.

Definition 20.2. *A process* X *is* **Adapted** *to the filtration* $\{\mathcal{F}_n\}_{n\in\mathcal{I}}$ *if* $\forall n \in \mathcal{I}$*, random variable* X_n *is* $\mathcal{F}_n$*-measurable.*

Definition 20.3. *A sequence of random variables* $S = \{S_n\}_{n\in\mathcal{I}}$ *is a* **Martingale** *with respect to* $\{\mathcal{F}_n\}_{n\in\mathcal{I}}$ *if*

(1) S *is adapted;*

(2) $\mathbb{E}[|\mathcal{F}_n|] < \infty \qquad \forall n \in \mathcal{I}$

(3) $\mathbb{E}[S_{n+1}|\mathcal{F}_n] = S_n \quad a.s. \quad \forall n \in \mathcal{I}$

Note that in condition (3), $\mathbb{E}[S_{n+1}|\mathcal{F}_n]$ is a random variable, therefore it equals to another random variable only in an "almost-surely" sense. The intuition of this property is that current knowledge is insufficient to predict the future any better than the present.

Definition 20.4. S *is called a* **Super-Martingale** *(and respectively* **Sub-Martingale***) if it satisfies condition (1),(2), and*

(4) $\mathbb{E}[S_{n+1}|\mathcal{F}_n] \leq (\geq) S_n \quad a.s. \quad \forall n \in \mathcal{I}$

20.3 The Doob-Kolmogorov Inequality

The Doob-Kolmogorov Inequality bounds the deviations of a stochastic process. This section starts from lemmas and incrementally builds the proof of this inequality.

Lemma 20.1. $\mathbb{E}[Xf(Y)|Y] = f(Y)\mathbb{E}[X|Y]$.

This is obvious because once Y is given, $f(Y)$ is deterministic and can be moved out of the expectation.

Lemma 20.2. $\mathbb{E}[\mathbb{E}[X|Y_{1:2}]|Y_1] = \mathbb{E}[X|Y_1]$. *This is known as the* **Tower Property**.

Proof:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X|Y_{1:2}]|Y_1 = y_1] \quad &= \int_X \int_{Y_2} x dP(X = x|Y_1 = y_1, Y_2) \\
&= \int_X x dP(X = x|Y_1 = y_1) = \mathbb{E}[X|Y_1 = y_1]
\end{aligned}$$

This is not surprising because unlike Y_1, Y_2 is non-deterministic and will be marginalized out. □

Lemma 20.3. *Assume a sequence* $\{S_n\}_{n\in\mathcal{I}}$ *is a martingale with respect to* $\{X_n\}_{n\in\mathcal{I}}$*, then* $\mathbb{E}[S_{m+n}|X_{1:m}] = S_m \qquad \forall n, m \in \mathcal{I}$.

Proof:

$$\begin{aligned}
\mathbb{E}[S_{m+n}|X_{1:m}] &= \mathbb{E}\left[\mathbb{E}[S_{m+n}|X_{1:m+n-1}]|X_{1:m}\right] \qquad \text{(by Lemma(20.2))} \\
&= \mathbb{E}[S_{m+n-1}|X_{1:m}] \\
&\vdots \qquad \text{(successive application of the Martingale property \textit{(3)})} \\
&= \mathbb{E}[S_{m+1}|X_{1:m}] \\
&= S_m
\end{aligned}$$

The intuition of the lemma is that if the best prediction of the immediate future is the present, so are the predictions of longer terms. □

Theorem 20.4. *If $S = \{S_n\}_{n\in\mathcal{I}}$ is a martingale with respect to $\{X_n\}_{n\in\mathcal{I}}$, then*

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

This is known as the **Doob-Kolmogorov Inequality**.

Proof: Let

$$A_n \triangleq \{|S_i| < \epsilon \quad \forall i \leq n\}$$
$$B_n \triangleq \{A_{n-1} \cap \{|S_n| \geq \epsilon\}$$

In other words, A_n is the set where none of the first n elements in S deviates from zero for more than ϵ, and B_n is the set where the first deviation occurs at the nth element. The centain event Ω can be written as:

$$\Omega = A_n \cup (\cup_{i=1}^n B_i)$$

Using the properties of conditional expecation and the fact that $P(\Omega) \equiv 1$:

$$\begin{aligned}
\mathbb{E}[S_n^2] &= P(\Omega)\mathbb{E}[S_n^2|\Omega] = \mathbb{E}\left[S_n^2\mathbb{I}_\Omega\right] \\
&= \mathbb{E}\left[S_n^2\mathbb{I}_{A_n\cup(\cup_{i=1}^n B_i)}\right] \\
&= \mathbb{E}\left[S_n^2\mathbb{I}_{A_n}\right] + \sum_{i=1}^n \mathbb{E}\left[S_n^2\mathbb{I}_{B_i}\right] \\
&\geq \sum_{i=1}^n \mathbb{E}\left[S_n^2\mathbb{I}_{B_i}\right] \qquad (20.1)
\end{aligned}$$

For each term in the summation:

$$\begin{aligned}
\mathbb{E}\left[S_n^2\mathbb{I}_{B_i}\right] &= \mathbb{E}\left[(S_n + S_i - S_i)^2\mathbb{I}_{B_i}\right] \\
&= \mathbb{E}\left[(S_n - S_i)^2\mathbb{I}_{B_i}\right] + 2\mathbb{E}\left[(S_n - S_i)S_i\mathbb{I}_{B_i}\right] + \mathbb{E}\left[S_i^2\mathbb{I}_{B_i}\right] \\
&= \alpha + 2\beta + \gamma \\
\alpha &\triangleq \mathbb{E}\left[(S_n - S_i)^2\mathbb{I}_{B_i}\right] \\
\beta &\triangleq \mathbb{E}\left[(S_n - S_i)S_i\mathbb{I}_{B_i}\right] \\
\gamma &\triangleq \mathbb{E}\left[S_i^2\mathbb{I}_{B_i}\right]
\end{aligned}$$

- The non-negativity of $(S_n - S_i)^2$ and $\mathbb{I}_{B_i}$ suggest that $\alpha \geq 0$;
- $|S_i| > \epsilon$ when B_i occurs, thus $\gamma \geq \epsilon^2 P(B_i)$

- $\beta = 0$ because:

$$\begin{aligned}\beta &= \mathbb{E}\left[\mathbb{E}[(S_n - S_i)S_i\mathbb{I}_{B_i}]|X_{1:i}\right] \\ &= \mathbb{E}\left[S_i\mathbb{I}_{B_i}\mathbb{E}[(S_n - S_i)|X_{1:i}]\right] \\ &= \mathbb{E}\left[S_i\mathbb{I}_{B_i}(S_i - S_i)\right] = 0 \qquad \text{(by Lemma(20.3))}\end{aligned}$$

Therefore $\mathbb{E}\left[S_n^2\mathbb{I}_{B_i}\right] \geq \epsilon^2 P(B_i)$. Substituting this into (20.1) gives:

$$\mathbb{E}[S_n^2] \geq \sum_{i=1}^{n} \epsilon^2 P(B_i) \geq \epsilon^2 P(\max_{1\leq i\leq n} |S_i| \geq \epsilon)$$

The last inequality results from the fact that $\{\max_{1\leq i\leq n} |S_i| \geq \epsilon\}$ is a subset of $\cup_{i=1}^n B_i$, i.e. if the maximum element deviates for more than ϵ, then at least one of the elements must have deviated for more than ϵ. This proves the Doob-Kolmogorov Inequality. □

20.4 The Big Picture

The diagram in figure 20.1 illustrates the connection between the proof of convergence for stochastic approximation (lecture 18) and the martigale theory (lecture 19 and this lecture).

In order to minimize the target function $f(\theta)$, a Lyapanov function (lecture 18) $\mathcal{L}(\theta)$ is constructed as a surrogate function to optimize. A stochastic approximation process is then executed to perform gradient updates of the Lyapanov function, yielding a sequence of parameters $\{\theta_n\}$.

As long as the process $\{S_n\}$ is a super-martingale with respect to the observation $\{F_n\}$, by definition, the sequence will converge to the optimal θ^* almost surely. Moreover, the probability that this process deviates beyond certain threshold is bounded, given by the Doob-Kolmogorov Inequality.

20.5 References

- David Williams, *Probability with Martingales*, Cambridge University Press, 1991, ISBN $0-521-40605-6$

- Geoffrey Grimmett and David Stirzaker, *Probability and Random Processes*, Oxford University Press, 2001, ISBM $978-0198572220$

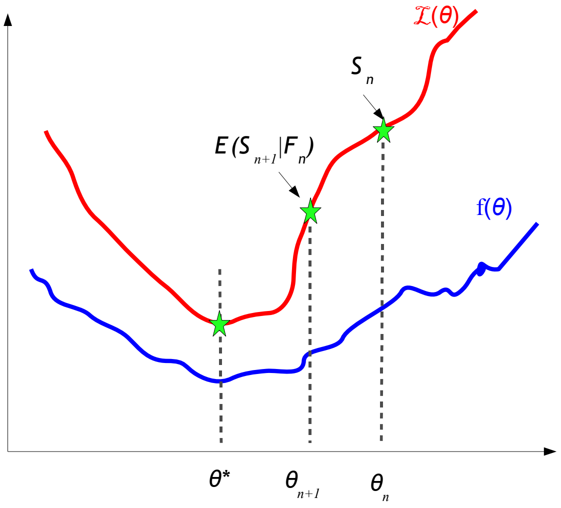

Figure 20.1. Stochastic Approximation.

- Wikipedia page on Martingales,
  http://en.wikipedia.org/wiki/Martingale_(probability_theory)
- Wikipeda page on Doob's Martingale Inequality,
  http://en.wikipedia.org/wiki/Doob's_martingale_inequality