CPSC 550: Machine Learning II	**2008/9 Term 2**
Lecture 18 — Mar. 26, 2009	
Lecturer: Nando de Freitas	*Scribe: Chris Nell*

This lecture begins a proof using Lyapunov functions and martingales of the convergence of stochastic approximation. The presentation here is similar to that in Chapter 4 of "Neuro-Dynamic Programming" by Bertsekas and Tsitsiklis [1].

18.1 Introduction

Recall from previous lectures the update equation for stochastic approximation:

$$\theta^{(i+1)} = \theta^{(i)} + \gamma^{(i+1)} \left(h(\theta^{(i)}) + \xi^{(i+1)} \right) \tag{18.1}$$

Here, $i \geq 0$ indexes the updates, $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ represents the parameters, γ is a learning rate, and $h(\theta)$ and ξ respectively represent the deterministic and stochastic components of the objective function. For example, in gradient descent of the parameter log-likelihood $\mathcal{L}(\theta)$:

$$\begin{aligned} h(\theta) &= \nabla_\theta \mathcal{L}(\theta) = g(\theta) = \mathbb{E}_f \left[g(\theta, X) \right] \\ \xi^{(i)} &= g(\theta) - g(\theta, X^{(i)}) \quad \text{where } X^{(i)} \sim f(X; \theta) \end{aligned}$$

However, note that h is not necessarily a gradient in general.

There are three distinct analysis methods by which Equation 18.1 can be proven to converge:

1. Lyapunov functions (using optimization for discrete settings, and martingales for stohastic settings)
2. Contraction operators; for example:
   - Markov chains produce contractions in L2, as each transition involves multiplication by a stochastic matrix with eigenvalues $\lambda \leq 1$
   - The Bellman operator
   - Bayes' rule is a contraction in KL-divergence of marginal likelihood, as additional data is added.
3. Differential equation analysis of $\theta(i)$

This lecture will present the first of these three methods, as among other reasons it is the method used by Younes in proving convergence of the maximum likelihood training procedure for Boltzmnn machines (see [2], or the notes for Lecture 12).

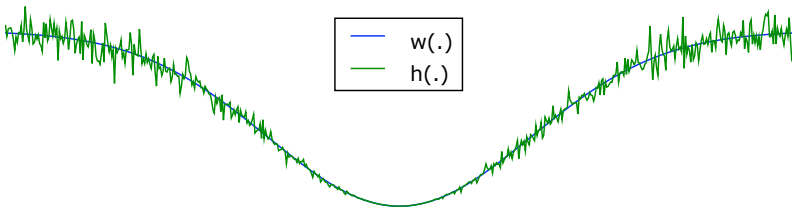

Figure 18.1. Illustration of h (noise) decreasing with w.

18.2 Lyapunov Functions

Let $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$

$w : \Theta \to \mathbb{R}$ (a scalar-valued function)

$h : \Theta \to \mathbb{R}^{d_\theta}$ (a vector-valued function)

Then w is a Lyapunov function with respect to (h, Θ) if:

- w is continuously differentiable on Θ
- $\langle \nabla_\theta w(\theta), h(\theta) \rangle \geq 0 \ \forall \theta$, with equality iff $h(\theta) = 0$

The inner product condition states that w and h "move together"; when one increases so does the other, as illustrated in Figure 18.1. In particular, if h represents noise, then the noise must decrease with w. This leads to the intuition behind the proof; showing that:[1]

$$\mathbb{E}\left[w\left(\theta^{(i)}\right) | \theta^{(1:i-1)}\right] \leq w\left(\theta^{(i-1)}\right)$$

The challenge in proving stochastic approximation to converge for a particular algorithm by this method is in finding a Lyapunov function for it. Sometimes this task is straightforward; for example, when the task is maximum likelihood estimation, an appropriate choice is:

$h(\theta) = g(\theta)$ (the gradient)

$w(\theta) = \mathcal{L}(\theta)$ (the log-likelihood)

[1] In this case, $w(\theta)$ is said to be a supermartingale.

18.3 Convergence of Stochastic Approximation

Theorem 18.1. Deterministic convergence [3, 4]

Assume that Θ is an open subset of $\mathbb{R}^{d_\theta}$, and let $h : \Theta \to \mathbb{R}^{d_\theta}$ be continuous. Let $\lim_{i\to\infty} \gamma^{(i)} = 0$, and let $\sum_{i=1}^{\infty} \gamma^{(i)} = \infty$. Let $\left\{\xi^{(i)}\right\}$ satisfy:

$$\lim_{k\to\infty} \sup_{L\geq k} \left| \sum_{i=k}^{L} \gamma^{(i)} \xi^{(i)} \right| = 0 \tag{18.2}$$

Assume further that there is a Lyapunov function w with respect to (h, Θ). Where $\Theta^* = \{\theta | h(\theta) = 0\}$ is the set of solutions, and $\theta^* \in \Theta^*$, assume that:

$$w(\theta^*) < \infty \tag{18.3}$$

Let $\theta^{(i)} = \theta^{(i-1)} + \gamma^{(i)} h\left(\theta^{(i-1)}\right) + \gamma^{(i)} \xi^{(i)}$ be such that $\theta^{(i)}$ remains in a compact space[2] *$K \subset \Theta$, with $\Theta^* \subset K$.*

If these conditions hold, then:

$$\lim_{i\to\infty} \theta^{(i)} = \theta^* \tag{18.4}$$

$$\lim_{i\to\infty} w\left(\theta^{(i)}\right) = w(\theta^*) = w^* \tag{18.5}$$

Proof: Omitted due to largely mechanical nature. □

Note that Theorem 18.1 requires that $\theta^{(i)} \in K \ \forall i$, where K is a compact space; this situation is illustrated in Figure 18.2. As in practice this assumption often fails, techniques for "forcing" compliance exist. For example, consider:

$$\theta^{(i+1)} = \Pi\left(\theta^{(i-1)} + \gamma^{(i)} h\left(\theta^{(i-1)}\right) + \gamma^{(i)} \xi^{(i)}\right) \tag{18.6}$$

where $\Pi : \Theta \to K$ is a projection operator. Alternatively, the approximation process can be restarted whenever some $\theta^{(i)} \notin K$ is encountered; if it can be shown that the number of restarts is finite, convergence still holds. [5, 6]

Also note that the treatment of Bertsekas and Tsitsiklis drops this boundedness requirement altogether, in exchange for assuming that $h(\theta) = g(\theta)$, a gradient. [1]

[2] A compact space is a bounded and closed space; $[1, 2]$ is compact, but $(1, 2)$ is not.

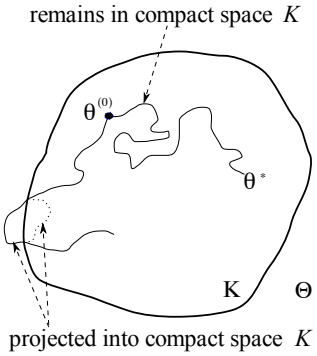

Figure 18.2. Illustration of two sequences of parameters $\{\theta^{(0:k)}\}$ and a compact space K. In one path, $\theta^{(i)} \in K \ \forall i$, and $\theta^{(k)} = \theta^*$, a solution, as required by Theorem 18.1. In the second, parts of the path leaving K ar projected back into K via Equation 18.6.

Theorem 18.2. Convergence of S.A.: ML for exponential families

Let $X^{(i)} \sim f(X;\theta)$, *where*

$$f(X;\theta) = \frac{h(X)}{Z(\theta)} \exp(\Psi(\theta)s(X)) \tag{18.7}$$

$$Z(\theta) = \int h(X) \exp(\Psi(\theta)s(X))dX \tag{18.8}$$

Assume $f(X,\theta)$ *is continuous on* Θ. *For* $r \geq 2$, *assume*

$$\int |s(X)|^r p(dX;\theta) < \infty$$

Choose $w(\theta) = \mathcal{L}(\theta) = \log f(X;\theta)$. *Assume:*

$$\gamma^{(i)} \geq 0 \quad \forall i$$

$$\sum_{i=1}^{\infty} \gamma^{(i)} = \infty$$

$$\sum_{i=1}^{\infty} {\gamma^{(i)}}^2 < \infty$$

Finally, assume that there exists some compact subset $K \subset \Theta$ *such that* $\theta^{(i)} \in K$ $\forall i$, *and also* $\theta^* \in K$ *for some solution* θ^*.

If these conditions hold, then almost surely $\theta^{(i)}$ *satisfies:*

$$\lim_{i \to \infty} \nabla_\theta \mathcal{L}(\theta^{(i)}) = 0 \tag{18.9}$$

Proof: In four parts; (C) and (D) to be continued next lecture.

(A)

$$\begin{aligned}\nabla_\theta \mathcal{L}(\theta) &= \nabla_\theta \Psi(\theta) s(X) - \nabla_\theta \log Z(\theta) \\ &= \nabla_\theta \Psi(\theta) s(X) - \frac{\nabla_\theta Z(\theta)}{Z(\theta)}\end{aligned}$$

Recalling Equation 18.8:

$$= \nabla_\theta \Psi(\theta) s(X) - \mathbb{E}\left[s(X)\right] \nabla_\theta \Psi(\theta) \tag{18.10}$$

Equation 18.10 can be interpreted as the difference between the sufficient statistics estimated at a point, and the expected sufficient statistics. Note that this is a generalization of what we did for RBMs (Equation 12.13 from Lecture 12), where in that case $\Psi(\theta) = I$, the identity matrix.

(B)

$$\begin{aligned}\xi^{(i)} &= \nabla_\theta \Psi\left(\theta^{(i-1)}\right)\left(-\left(s(X) - s\left(X^{(i)}\right)\right) + \left(s(X) - \mathbb{E}\left[s(X)\right]\right)\right) \\ &= \nabla_\theta \Psi\left(\theta^{(i-1)}\right)\left(s\left(X^{(i)}\right) - \mathbb{E}\left[s(X)\right]\right)\end{aligned} \tag{18.11}$$

Equation 18.11 can be interpreted as the difference between the gradient evaluated at data point $X^{(i)}$ and the expected gradient.

(C)

Define $M_j = \sum_{i=1}^{j} \gamma^{(i)} \xi^{(i)}$. Then if $M_j < \infty$, $\lim_{k\to\infty} \sup_{L \geq k} \left| \sum_{i=k}^{L} \gamma^{(i)} \xi^{(i)} \right| = 0$.

As will be demonstrated next lecture, it can be easily shown that:

$$\begin{cases} \mathbb{E}\left[M_j | X^{(1:j-1)}\right] = M_{j-1} \\ \sum_j \mathbb{E}\left[\left(M_j - M_{j-1}\right)^2 | X^{(1:j-1)}\right] &= \sum_j \mathbb{E}\left[\left(\gamma^{(j)} \xi^{(j)}\right)^2 | X^{(1:j-1)}\right] \\ &\leq \sum_j {\gamma^{(j)}}^2 ||\nabla_\theta \Psi\left(\theta^{(i-1)}\right)||^2 \mathbb{E}\left[s(X)^2\right] \end{cases}$$

where $||\nabla_\theta \Psi\left(\theta^{(i-1)}\right)||$ is finite due to compactness of K and differentiability of w, and $\mathbb{E}\left[s(X)^2\right]$ is finite by assumption of finite moments; thus:

$$\sum_j \mathbb{E}\left[(M_j - M_{j-1})^2 \,|X^{(1:j-1)}\right] < \infty \tag{18.12}$$

(D)

It will be shown that since M_j is a martingale, $M_\infty < \infty$. Thus, so long as samples $X^{(i)}$ are i.i.d., this allows proof of convergence for for any exponential family model (such as RBMs). Techniques to relax the i.i.d. assumption (for example, when sampling from a Markov chain) will also be discussed.

□

For more information on martingales, the book of Williams is recommended. [7]

Bibliography

[1] Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-dynamic Programming. Athena Scientific. (1996)

[2] Younes, L.: Parameter estimation for imperfectly observed Gibbsian fields. Prob. Theory and Rel. fields 82, 625–645 (1989)

`http://cis.jhu.edu/ younes/Preprints/younes.MRFpartial.pdf`

[3] Andrieu, C., Moulines, E., Priouret, P.: Stability of stochastic approxi- mation under veriable conditions. Siam Journal On Control and Optimization, 44, 283312. (2005)

[4] Cappé, C., Moulines, E., Rydén, T.: Inference in Hidden Markov Models. Springer Series in Statistics. (2005)

[5] V. Tadić: Asymptotic analysis of stochastic approximation algorithms under violated Kushner-Clark conditions with applications. CDC 2000.

[6] Meyn, S., and Tweedie, R.: Markov Chains and Stochastic Stability. Communications and Control Engineering Series. Springer-Verlag. (1993)

`http://probability.ca/MT/`

[7] Williams, D.: Probability with Martingales. Cambridge Mathematical Textbooks. (1991)