Stochastic Simulation: Lecture 16

Prof. Mike Giles

Oxford University Mathematical Institute

Objectives

The stochastic approximation problem is to determine θ such that

$$\mathbb{E}[f(\theta, X)] = 0,$$

If $f = \nabla V$ then this also corresponding to minimising or maximising

$$\mathbb{E}[V(\theta, X)].$$

In Machine Learning this corresponds to maximising the log-likelihood given a large set of data:

$$\text{log-likelihood} \;=\; \sum_{i=1}^{S} L_i(\theta) \;=\; \mathbb{E}[S\, L_I(\theta)]$$

where the expectation comes from taking a random index I, uniformly distributed over $\{1, 2, \ldots, S\}$.

Steepest descent

The classic steepest descent method for solving $f(\theta) = 0$ is based on a time-discretisation of

$$\dot{\theta} = -f(\theta)$$

which gives

$$\theta_{n+1} = \theta_n - k\, f(\theta_n).$$

From this we get

$$\theta_{n+1} - \theta_n = (I - kJ)\,(\theta_n - \theta_{n-1})$$

where $J \equiv \partial f / \partial \theta$.

So it converges to the root θ^* from near θ^* if $\|I - kJ\| < 1$.

Robbins-Munro

Starting from

$$\theta_{n+1} = \theta_n - k\,\mathbb{E}[f(\theta_n, X)]$$

the idea of Robbins & Munro was to replace the expectation by a single sample to give

$$\theta_{n+1} = \theta_n - k_n\, f(\theta_n, X_n)$$

with independent samples X_n.

If we write $F(\theta) \equiv \mathbb{E}[f(\theta, X)]$ then we can write this as

$$\theta_{n+1} = \theta_n - k_n\, F(\theta_n) - k_n\,(f(\theta_n, X_n) - F(\theta_n))$$

Robbins-Munro

Consider now the SDE

$$\mathrm{d}\theta_t = -F(\theta_t)\,\mathrm{d}t + \sigma(\theta_t)\,\mathrm{d}W_t$$

which has discretisation with timestep k_n

$$\theta_{n+1} = \theta_n - k_n\,F(\theta_n) + \sigma\sqrt{k_n}\,Z_n$$

Equating this (approximately) to

$$\theta_{n+1} = \theta_n - k_n\,F(\theta_n) - k_n\,(f(\theta_n, X_n) - F(\theta_n))$$

gives

$$\sigma^2 \;\approx\; k_n\,\mathbb{V}[f(\theta_n, X_n)]$$

Conclusion? For convergence we need $\displaystyle\sum_n^{\infty} k_n \to \infty, \quad k_n \to 0$

Robbins-Munro

Usually, the second condition is tightened to $\sum_{n}^{\infty} k_n^2 < \infty$.

A frequent choice is $k_n = a/n$.

After running the iteration for N steps, the output of the Robbins-Munro algorithm is the final value θ_N.

Polyak and Ruppert independently improved this by using an average for the output

$$\overline{\theta}_N \equiv N^{-1} \sum_{1}^{N} \theta_n$$

– the averaging cancels out a lot of the noise in θ_n

Stochastic Gradient method

Similarly, the stochastic gradient method

$$\theta_{n+1} = \theta_n - k_n \nabla L_{I_n}(\theta_n)$$

where $L_i(\theta)$ is the log-likelihood associated with the i^{th} data item, and I_n is the random data index on step n.

An alternative is to use a mini-batch of samples in step n:

$$\theta_{n+1} = \theta_n - k_n \frac{1}{m} \sum_{1}^{m} \nabla L_{I_{n,m}}(\theta_n)$$

– no mathematical benefit, but provides scope for parallelisation or vectorisation.

(Practical question: do they use sampling with or without replacement?)

Stochastic Gradient method

There are lots of different variants of the stochastic gradient method, with the objective of achieving faster convergence.

One major line of development incorporates "momentum".

In the simplest form, this involves adding in a multiple of the previous correction:

$$\theta_{n+1} = \theta_n - k_n \nabla L_{I_n}(\theta_n) + \alpha_n(\theta_n - \theta_{n-1})$$

with $0 < \alpha_n < 1$.

Stochastic Gradient Lagrangian Dynamics (SGLD)

Another variant is to add in additional noise, to approximate the SDE

$$\mathrm{d}\theta_t = -\,\mathbb{E}[\nabla L_I(\theta_n)]\,\mathrm{d}t + \sigma\,\mathrm{d}W_t$$

using

$$\theta_{n+1} = \theta_n - k\,\nabla L_{I_n}(\theta_n) + \sigma\sqrt{k}\,Z_n$$

where the Z_n are iid $N(0,1)$ r.v.'s.

Lukas Szpruch has looked at using MLMC for this.

Key references

https://en.wikipedia.org/wiki/Stochastic_approximation

https://en.wikipedia.org/wiki/Stochastic_gradient_descent

H. Robbins, S. Monro. "A Stochastic Approximation Method". The Annals of Mathematical Statistics. 22(3):400, 1951

B.T. Polyak, A.B. Juditsky. "Acceleration of Stochastic Approximation by Averaging". SIAM Journal on Control and Optimization. 30(4):838, 1992

Key references

N. Frikha. "Multi-level stochastic approximation algorithms". Annals of Applied Probability, 26(2):933-985, 2016.

S. Dereich, T. Müller-Gronbach. "General multilevel adaptations for stochastic approximation algorithms of Robbins-Munro and Polyak-Ruppert type". Numerische Mathematik, 142(2):279-328, 2019.

D.E. Rumelhart, G.E. Hinton, R.J. Williams. 'Learning representations by back-propagating errors'. Nature. 323(6088):533-536, 1986.

M.B. Giles, M.B. Majka, L. Szpruch, S. Vollmer, K. Zygalakis. 'Multilevel Monte Carlo methods for the approximation of invariant measures of stochastic differential equations', Statistics and Computing, 30(3):507-524, 2020.