

## A Stochastic Approximation Method (Robbins et al., 1951)

Satvik Saha (ss7228)

September 23, 2025

Robbins and Monro [1951] describes a root finding algorithm which replaces direct computation of the function with a (random) estimate. Given a suitable increasing function  $M$ , one can solve for its root  $x^*$  (satisfying  $M(x^*) = 0$ ) by iterating  $x_{n+1} = x_n - a_n M(x_n)$  with an appropriate step size sequence  $\{a_n\}$ . Assuming that  $M$  is differentiable with  $0 < \eta < M' < \gamma$  gives  $M(x_n) - M(x^*) = M'(x'_n)(x_n - x^*)$  for some  $x'_n$  between  $x_n, x^*$ , hence with all  $0 < a_n \leq 1/\gamma$ ,

$$\begin{aligned} |x_{n+1} - x^*| &= |x_n - x^* - a_n M'(x'_n)(x_n - x^*)| \\ &= (1 - a_n M'(x'_n)) |x_n - x^*| \\ &\leq (1 - \eta a_n) |x_n - x^*| \\ &\leq e^{-\eta a_n} |x_n - x^*| \\ &\leq e^{-\eta \sum_{k=1}^n a_k} |x_1 - x^*|. \end{aligned}$$

If  $\{a_n\}$  is chosen such that  $\sum_{n=1}^{\infty} a_n = \infty$  (optimally,  $a_n = 1/\gamma$ ), we have convergence with  $x_n \rightarrow x^*$ .

If computing  $M(x)$  is intractable, we may instead use an estimate  $y(x)$ , specifically an unbiased estimator satisfying  $\mathbb{E}[y(x) | x] = M(x)$  and  $\mathbb{E}[y(x)^2 | x]$  uniformly bounded. Our new iterative scheme is  $x_{n+1} = x_n - a_n y_n$ , where  $y_n$  is an instance of  $y(x_n)$ . With some careful conditions on  $M$  and  $y$ , Robbins and Monro [1951] show that for sequences  $\{a_n\}$  of type  $1/n$  (we want  $c/n < a_n < c'/n$  for constants  $c, c' > 0$ ), we have convergence with  $x_n \rightarrow x^*$  in  $L^2$  (hence in probability).

In the context of minimizing a function  $f$ , the gradient  $\nabla f$  plays the role of  $M$  here; the algorithm is one of *stochastic gradient descent* (SGD). The conditions on  $M$  thus correspond to assumptions on the smoothness and convexity of  $f$ .

We have seen how SGD can be used effectively to compute quantities like *maximum a posteriori* (MAP) estimates of parameters in probabilistic models. Here, the gradient of the objective function to be minimized (the negative log joint density) is often expensive to compute since it depends on all of the given data  $x$ . Since an unbiased estimate of  $\frac{1}{n} \sum_{i=1}^n g(x_i)$  is  $g(x_I)$  (or indeed  $\frac{1}{m} \sum_{j=1}^m g(x_{I_j})$ ), where  $I$  (or each  $I_j$ ) is uniformly random over the indices  $\{1, \dots, n\}$ , SGD lets us deal with gradients of this form using a small fraction of our data at each iterate.

Stochastic gradients are a key idea in Ranganath et al. [2014]: the gradient of the ELBO to be optimized is an expectation over the variational distribution  $q$ , which is estimated by a Monte-Carlo average using samples from  $q$ .

## References

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL <https://doi.org/10.1214/aoms/1177729586>.