MEASURE VALUED DIFFERENTIATION FOR STOCHASTIC NEURAL NETWORKS

Thomas Flynn

Department of Computer Science
The Graduate Center, CUNY
365 5th Ave.
New York, NY 10016, USA

ABSTRACT

Stochastic neural networks serve as general models useful for machine learning problems. Several models on discrete state spaces have been studied, and their proposed gradient estimation procedures are based on having closed form solutions for the resulting probability distributions. The methods exploit constraints on network connectivity, such as symmetry, or the absence of cycles. Our interest is in the general case of long-term average cost in networks with arbitrary connectivity, where only knowledge of the transition probabilities is available. We propose an algorithm that computes descent directions based on simultaneous perturbation analysis and measure valued differentiation.

1 BACKGROUND

The earliest neural network models to be studied from the computational perspective were the deterministic threshold networks (Rosenblatt 1958). Unfortunately, their non-smooth nature prevents these models from being used in complex problems such as image recognition. One solution is to randomize the network using noise that is in some way smooth, and perform gradient based optimization that exploits smoothness of the resulting probabilities. A very general model for such a network was introduced in (Little 1974).

The Little model can be interpreted as a threshold network, where the thresholds are randomly chosen at each time step. A formal definition can be given as follows. Let the network have n nodes, and let $\xi(1), \xi(2)\dots$ be a sequence of noise vectors in $\mathbb{R}^n$, with the entire collection $\{\xi_i(t); i=1,\dots,n, t=1,2,\dots\}$ independent and distributed according to the logistic distribution. That is, the CDF of $\xi_i(t)$ is $P(\xi_i(t) < x) = \sigma(x) = \frac{1}{1+\exp(-x)}$. Define $f : \{0,1\}^n \times \Theta \times \Xi \to \{0,1\}^n$ in terms of the component functions $f_1,\dots,f_n$ as

$$f_i(x,(w,b),\xi) = \begin{cases} 1 \text{ if } \sum\limits_{j=1}^{n} w_{i,j}x_j + b_i > \xi_i \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

This function f and the noise $\xi(1), \xi(2)\dots$ determines the operation of the network. From the initial point $x(0)$, generate each successive point $x(n)$ following the recursion

$$x(t+1) = f(x(t), \theta, \xi(t+1)). \tag{2}$$

We denote by $P_\theta(x^0, x^1)$ be the probability of going to state x^1 from state x^0.

The equation (2) defines an ergodic Markov chain. Let d_{TV} be the total variation metric. It can be shown it possesses a stationary measure π_θ for any θ, and, setting $\varepsilon = \sigma(-\|w\|_\infty - \|b\|_\infty)^n$, for any initial measure μ it holds that

$$d_{TV}(\mu P_\theta^k, \pi_\theta) \leq (1-\varepsilon)^k d_{TV}(\mu, \pi_\theta).$$

Given the stationary measure π, and a cost function $e : \{0,1\}^n \to \mathbb{R}$, we define the overall objective:

$$J(\theta) = \mathbb{E}_{x\sim\pi_\theta}[e(x)]$$

978-1-5386-3428-8/17/$31.00 ©2017 IEEE

To compute the derivative of J, we will not attempt to compute the stationary the distribution. Rather we are interested in a gradient estimator which uses only knowledge of the kernel P.

2 RESEARCH QUESTION

The idea of measure valued differentiation is to express the derivative of an expectation as the scaled difference of two expectations (Heidergott, Vázquez-Abad, Pflug, and Farenhorst-Yuan 2010). If these measures are easy to sample from, this leads to a simple, unbiased derivative estimator. The concept of MVD can be extended from measures to Markov kernels, and then applied to derivatives of stationary costs (Pflug 1992).

In general for measure valued differentiation, m simulations are required for a system with m parameters. To address this potentially burdensome requirement, we propose a gradient estimator that works by picking a random direction, as in the simultaneous perturbation approach of (Spall 1992), and then computes the directional derivative using measure valued differentiation. The method is termed simultaneous perturbation measure valued differentiation. The only requirement is that one can compute the measure valued derivative along arbitrary directions.

Let μ_θ be a measure depending on an n-dimensional vector parameter θ. Let $v \in \mathbb{R}^n$ be a direction. A triple $(c_{\theta,v}, \mu^+_{\theta,v}, \mu^-_{\theta,v})$ is called a *measure valued directional directional derivative in the direction v at* θ if for all $e : X \to \mathbb{R}$,

$$\tfrac{\partial}{\partial \theta} \mu_\theta(e) v = c_{\theta,v}[\mu^+_{\theta,v}(e) - \mu^-_{\theta,v}(e)].$$

Defining the extension to Markov chains is straightforward. A triple $(c_{\theta,v}, P^+_{\theta,v}, P^-_{\theta,v})$ is a measure valued derivative for the Markov kernel P in the direction v if for each x, $(c_{\theta,v}(x), P^+_{\theta,v}(x,\cdot), P^-_{\theta,v}(x,\cdot))$ is an MVD in the direction v for the measure $P_\theta(x,\cdot)$.

The proposed gradient estimator for stationary costs proceeds by choosing a random direction, and applying the stationary MVD algorithm of (Pflug 1992). The estimator will be applied to our motivating example, the Little model. The theoretical questions involve relating the variance and convergence properties of the estimator to the properties of the underlying neural network. On the practical side, we are interested in using the estimator to apply the Little networks to machine learning tasks such as image classification.

REFERENCES

Heidergott, B., F. J. Vázquez-Abad, G. Pflug, and T. Farenhorst-Yuan. 2010, February. “Gradient Estimation for Discrete-event Systems by Measure-valued Differentiation”. *ACM Trans. Model. Comput. Simul.* 20 (1): 5:1–5:28.

Little, W. A. 1974. “The existence of persistent states in the brain”. *Mathematical biosciences* 19 (1-2): 101–120.

Pflug, G. C. 1992. “Gradient estimates for the performance of markov chains and discrete event processes”. *Annals of Operations Research* 39 (1): 173–194.

Rosenblatt, F. 1958. “The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain”. *Psychological Review*:65–386.

Spall, J. C. 1992. “Multivariate stochastic approximation using a simultaneous perturbation gradient approximation”. *IEEE transactions on automatic control* 37 (3): 332–341.