Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physics Letters A 350 (2006) 81–86
PHYSICS LETTERS A
www.elsevier.com/locate/pla

Geometry on the parameter space of the belief propagation algorithm on Bayesian networks ☆

Yodai Watanabe

National Institute of Informatics, Research Organization of Information and Systems, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

Received 25 July 2003; received in revised form 20 June 2005; accepted 5 October 2005

Available online 11 October 2005

Communicated by A.P. Fordy

Abstract

This Letter considers a geometrical structure on the parameter space of the belief propagation algorithm on Bayesian networks. The statistical manifold of posterior distributions is introduced, and the expression for the information metric on the manifold is derived. The expression is used to construct a cost function which can be regarded as a measure of the distance in the parameter space.

PACS: 89.70.+c; 02.50.-r; 02.40.Ky; 02.60.Pn

Keywords: Statistical manifold; Belief propagation algorithm; Bayesian networks

1. Introduction

Bayesian networks are graphical representations of probabilistic dependence among random variables. For a given Bayesian network, probabilistic inference in the network is to evaluate the posterior distribution (or belief) $P(X = x|\boldsymbol{E} = \boldsymbol{e})$, where X is a random variable associated with the network, x is an assignment of X, and the event $\boldsymbol{E} = \boldsymbol{e}$ is an observed evidence. It is known that the exact probabilistic inference for general Bayesian networks is NP-hard [3], while there exist efficient inference algorithms for a few special classes of Bayesian networks. One of the most successful applications of such algorithms can be found in the field of the error-correcting codes.

The error-correcting codes are fundamental techniques for improving the reliability of communication over a noisy transmission channel, and have many applications in various communication and data storage systems. The idea of error correction is the introduction of redundancy, which makes it possible to detect or remove errors in the received information. It is clear that one can transmit information with an arbitrarily small probability of error by simply increasing the redundancy to infinity (or equivalently, by decreasing the transmission rate to zero). However, it is known that an arbitrarily small probability of error is achievable with the transmission rate kept finite; in fact, the channel coding theorem states that it is possible to transmit information at any rate below channel capacity with a probability of error arbitrarily close to zero, where channel capacity is a real-valued parameter which characterizes the channel [16]. Since the appearance of this theorem, a great deal of research has been devoted to the problem of designing efficient error-correcting codes which achieve channel capacity.

Recently, as the most promising solutions to the problem, the turbo codes were invented in [2], and the low-density parity-check (LDPC) codes, originally proposed in [4], were rediscovered in [11]. Since these codes have attractive features such as long length of codes, randomness in encoding and approximation in decoding, there has been much research based on various fields such as artificial intelligence [9,12], statistical physics [7,8] and dynamical systems [15]. So far, it has been shown experimentally that these codes have excellent performance approaching channel capacity, while the theoretical understanding

☆ Research supported in part by the Special Postdoctoral Researchers Program of RIKEN (The Institute of Physical and Chemical Research).

0375-9601/$ – see front matter
doi:10.1016/j.physleta.2005.10.012

is still weak. One of the most important clues to the understanding would be the fact that the decoding algorithms of these codes are equivalent to the belief propagation (BP) algorithm [9,12], which is an exact inference algorithm for Bayesian networks with no loops [13]. Although this result is important and suggestive, it is not guaranteed that the BP algorithm will perform valid computation on the networks of these code because they have loops. Thus the dynamics of the BP algorithm on networks with loops has been investigated in order to examine the connection between performance of the algorithm and structure of the corresponding networks [18,19].

In studying or designing inference algorithms, it may often be useful and suggestive to take into account a geometrical structure of the space of the inference parameters. For instance, the dynamics of the turbo decoding can be intuitively described in terms of information geometry, which may lead to an insight into designing the codes with better performance [6]. Also, in the field of the statistical learning, it has been reported that the natural gradient descent, an inference algorithm which takes advantage of the Riemannian structure on the parameter space induced by the information metric, shows much better convergence than the conventional inference algorithm (the steepest gradient descent) [14].

In this Letter, we consider geometry on the space of the inference parameters of the BP algorithm. For this purpose, we first introduce the statistical manifold of the posterior distributions (or beliefs), and derive the expression for the information metric on the manifold in terms of the inference parameters of the BP algorithm. By using the expression, we construct a cost function which can be used as a measure of the distance in the parameter space. Further, by use of the cost function, we provide an inference algorithm which has a superlinear rate of convergence in the space of the posterior distributions. See, e.g., [10] for general discussions on the rate of convergence of algorithms in optimization.

2. Probabilistic inference in Bayesian networks

We start by providing a brief introduction to the probabilistic inference in Bayesian networks. (For details, see, e.g., [13].) A Bayesian network is a directed acyclic graph (DAG) which represents probabilistic dependence among random variables. More precisely, a Bayesian network is a triplet $B = \langle V, E, P\rangle$ such that

(1) V is a set of random variables: $V = \{V_1, \ldots, V_N\}$;
(2) $\langle V, E\rangle$ is a DAG;
(3) P is a set of conditional probabilities: $P = \{P(v_i|pa(v_i)) \mid V_i \in V\}$, where v_i and $pa(v_i)$ denote value assignments for V_i and parents of V_i, respectively;
(4) The joint distribution $P(v_1, \ldots, v_N)$ is factored according to structure of the graph $\langle V, E\rangle$ in the form

$$P(v_1, \ldots, v_N) = \prod_{i=1}^{N} P\big(v_i|pa(v_i)\big). \tag{1}$$

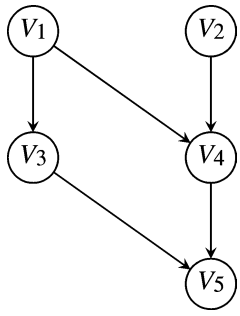

Fig. 1. Simple example of a DAG of 5 nodes and 5 edges.

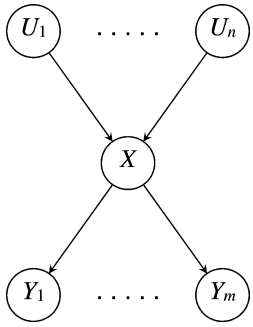

Fig. 2. Local structure of a graph.

Fig. 1 illustrates a simple example of a DAG. In this case, for example, the joint probability distribution $P(v_1, v_2, v_3, v_4, v_5)$ is written as

$$P(v_1, v_2, v_3, v_4, v_5) = P(v_1)P(v_2)P(v_3|v_1)P(v_4|v_1, v_2)P(v_5|v_3, v_4). \tag{2}$$

Now, let $\boldsymbol{E}$ be a subset of V, and suppose that the evidence $\boldsymbol{E} = \boldsymbol{e}$ is observed. The probabilistic inference in a Bayesian network is then to compute the posterior distribution (or belief) $BEL_{V_i}(v_i) = P(v_i|\boldsymbol{e})$ for all $V_i \notin \boldsymbol{E}$.

One straightforward approach to computing the belief $BEL_{V_i}(v_i)$ is to take the sum of all the possible terms of $P(v_1, \ldots, v_N)$,

$$BEL_{V_i}(v_i) = \alpha \sum_{v_j \notin \{v_i, \boldsymbol{E}\}} P(v_1, \ldots, v_N), \tag{3}$$

where α is the normalization constant. However, this approach requires exponential number of operations with respect to the size of the system (i.e., the number of unobserved random variables), and so is impractical unless the size is not large.

In order to describe a more practical approach, let us take the case when a graph has no loops, and turn our consideration to a local structure of the graph, namely, a node $X \notin \boldsymbol{E}$, its parents $\boldsymbol{U} = \{U_1, \ldots, U_n\}$ and its children $\boldsymbol{Y} = \{Y_1, \ldots, Y_m\}$ (see Fig. 2). Since the graph has no loops, the path from an evidence node $E \in \boldsymbol{E}$ to X is uniquely determined, and hence $\boldsymbol{E}$ is separable as $\boldsymbol{E} = \{\boldsymbol{E}^+_{U_iX}, \ldots, \boldsymbol{E}^-_{XY_j}, \ldots\}$ according to the intersection of the path with the neighbouring nodes $\{\boldsymbol{U}, \boldsymbol{Y}\}$. Taking this separation into account, we now introduce the parameters $\pi_{U_iX}(u_i)$ and $\lambda_{XY_j}(x)$ by writing

$$\pi_{U_iX}(u_i) = P\big(u_i|\boldsymbol{e}^+_{U_iX}\big), \qquad \lambda_{XY_j}(x) = P\big(\boldsymbol{e}^-_{XY_j}|x\big), \tag{4}$$