Variational Maximum A Posteriori by Annealed Mean Field Analysis

Gang Hua, *Student Member*, *IEEE*, and Ying Wu, *Member*, *IEEE*

Abstract—This paper proposes a novel probabilistic variational method with deterministic annealing for the maximum a posteriori (MAP) estimation of complex stochastic systems. Since the MAP estimation involves global optimization, in general, it is very difficult to achieve. Therefore, most probabilistic inference algorithms are only able to achieve either the exact or the approximate posterior distributions. Our method constrains the mean field variational distribution to be multivariate Gaussian. Then, a deterministic annealing scheme is nicely incorporated into the mean field fix-point iterations to obtain the optimal MAP estimate. This is based on the observation that when the covariance of the variational Gaussian distribution approaches to zero, the infimum point of the Kullback-Leibler (KL) divergence between the variational Gaussian and the real posterior will be the same as the supreme point of the real posterior. Although global optimality may not be guaranteed, our extensive synthetic and real experiments demonstrate the effectiveness and efficiency of the proposed method.

Index Terms—Mean field variational analysis, deterministic annealing, maximum a posteriori estimation, graphical model, Markov network.

1 Introduction

Bayesian inference methods recover the posterior distribution $P(\mathbf{X}|\mathbf{Z})$, or find the maximum a posteriori (MAP) estimation $\hat{X} = \arg\max_X\{P(\mathbf{X}|\mathbf{Z})\}$, where $\mathbf{Z}$ is the set of observations of the stochastic system and $\mathbf{X}$ is the underlying stochastic processes generating $\mathbf{Z}$. Many real problems can be effectively modeled and solved under the Bayesian inference framework. In the literature of signal processing and computer vision, Bayesian methods are widely used in signal estimation [1], image segmentation [2], [3], image super-resolution [4], [5], and visual tracking [6], [7], [8], etc. Many of these Bayesian inference problems are formulated and represented by probabilistic graphical models [4], [5], [6], [7], [8].

Most traditional methods of Bayesian inference such as the belief propagation (BP) algorithm [4], [5], [9] and the variational inference methods [7], [10], [11], [12] focus on recovering either the exact or the approximate posterior distributions. The problem is that even if we could obtain the exact posterior distribution, in general, it is still very difficult to find the MAP estimate since it involves global optimization. The Markov chain Monte Carlo (MCMC) technique with simulated annealing (SA) [13], [14], [15] provides a principled way to search for the global optimum of the posterior and the convergence in probability to the global optimum has been proven [15]. However, the SA schemes are usually computationally intensive, which hinders their applicability in many real applications.

In this paper, we propose an efficient approach to finding the MAP estimate by an annealed mean field variational analysis. We show that when the covariance of the variational Gaussian distribution approaches to zero, the infimum point of the KL divergence between the variational Gaussian and the real posterior will be the same as the supreme point of the real posterior. Thus, in the limit, minimizing the KL divergence between the variational Gaussian and the real posterior is equivalent to maximizing the real posterior. The advantage of minimizing the former is that we can nicely incorporate a deterministic annealing (DA) scheme [16], [17], [18], [19] into the mean field fix-point iterations, which will eventually converge into the optimal or a near-optimal maximum point of the real posterior. This new method, namely, variational MAP, is an efficient and effective method for obtaining the MAP estimate of a complex stochastic system.

The remainder of this paper is organized as follows: In Section 2, related work are categorized and discussed. Then, in Section 3, we construct the theoretic foundation of the variational MAP algorithm by revealing a general theorem of the KL divergence between a Gaussian and an arbitrary p.d.f. In Section 4, without loss of generality, we deduce the mean field fix-point iterations under a Markov network, where the mean field approximation is constrained to be a multivariate Gaussian. We then propose the variational MAP algorithm in Section 5. Furthermore, a Monte Carlo implementation of such a variational MAP algorithm is proposed in Section 6. Extensive experimental results are demonstrated and discussed in Section 7. Finally, we conclude our work and propose the possible future work in Section 8.

2 Related Work

We propose the variational MAP algorithm under the context of graphical model since it is a powerful means of representing real stochastic systems. Moreover, the MAP estimate involves global optimization. Related work can thus be categorized into three. The first category is related to graphical model representation of stochastic systems. The

• *The authors are with the Department of Electrical and Computer Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208. E-mail: {ganghua, yingwu}@ece.northwestern.edu.*

Manuscript received 10 June 2004; revised 21 Mar. 2005; accepted 28 Mar. 2005; published online 14 Sept. 2005.
Recommended for acceptance by Y. Amit.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0296-0604.

second category involves the Bayesian inference algorithms on graphical models, while the third category is related to the global optimization methods.

Bayesian network (BN), dynamic Bayesian network (DBN) [20], [21], Markov network [4], [5], and dynamic Markov network [7], [10], [22] are all typical graphical models [23]. They are widely used for modeling and solving computer vision problems. To mention some, a BN is proposed in [24] for spatial-temporal segmentation of video sequences. Various DBNs are proposed to address different problems in visual tracking, such as multiple cue coinference [6], switching observation models for contour tracking in clutter [25], and tracking the appearances of multiple targets against occlusion [26]. The Markov network is adopted to achieve image super-resolution [4], [5], while various dynamic Markov networks are adopted to perform articulated human body tracking [7], to analyze structured deformable shapes [10], and to formulate a rigorous bidirectional multiscale visual tracking algorithm to address the abrupt motion [22]. Although there are many types of graphical models, they all can be transformed into one another [23].

For Bayesian inference in graphical models, when there is no loop, the sum-product algorithm or belief propagation (BP) [23], [4] can obtain the exact inference efficiently through a local message passing process. When there are loops, the loopy BP [27] and generalized BP [9] can obtain good approximate results [4], [28]. As an approximation, Monte Carlo techniques such as Markov chain Monte Carlo (MCMC) [23], [2], [3] and sequential Monte Carlo [29], [30], [31] can be used for implementing the Bayesian inference by sampling. In addition, probabilistic variational approach provides a principled way for approximate inference such as the mean field variational analysis [12], [11], [32], [7], [10], which seeks the best approximate results by minimizing the KL divergence between the mean field approximation and the real posterior distribution.

The nonparametric BP [33] and the PAMPAS algorithm [34] are proposed to implement the Bayesian inference on complex real valued graphical models by combining the BP algorithm with the MCMC sampler. A different approach is the sequential mean field Monte Carlo algorithm (SMFMC) [7], [10], which combines the mean field variational analysis with the sequential Monte Carlo technique. It is also proposed to implement efficient Bayesian inference on complex real valued graphical models.

Finding the MAP estimate is a global optimization problem. In terms of complexity, it is a NP-hard problem in the combinatory context. However, the stochastic simulated annealing (SA) [13], [15], [14] can achieve good results in many applications since the convergence in probability to the global optimum is proven [15]. But, SA algorithms are often inherently slow due to their randomized local search strategy. Deterministic annealing (DA) [16], [17] methods intend to overcome the inefficiency of the SA methods. They are based on deterministic optimization scheme, but they incorporate stochastic smoothing by optimizing over a probabilistic state space [17]. Although global optimality may not be guaranteed for DA, many empirical studies have shown that the DA methods are very likely to achieve optimal or near optimal solutions [17]. The annealing methods are enlightened by the annealing process of a thermodynamic system, which drives the system to stay in the lowest energy and, thus, most probable state. Annealing methods have been widely used in image processing, computer vision, and pattern recognition for robust M-Estimation [19], for designing piecewise regression models [35], for image texture segmentation and grouping [18], and for object recognition [36], to list a few.

In [37], an annealed particle filtering algorithm, which integrates a SA scheme with the sequential Monte Carlo algorithm, is proposed to find the maximum of the articulated human motion posteriors. Instead of using MCMC, weighted resampling is preformed during the SA process. Notwithstanding the empirically demonstrated effectiveness, this algorithm is largely based on heuristics and there is no strict theoretic proof about the convergence of such a process.

The variational MAP algorithm proposed in this paper integrates the mean field variational inference method [23], [7], [10], [12], [11] with a DA scheme [16], [17], [18]. By constraining the mean field variational distribution to be a multivariate Gaussian, the covariance of the Gaussian will be used as the "temperature" for annealing. And, in each step of the annealing, we iterate the Gaussian mean field fix-point equations to converge. As the covariance of the variational Gaussian approaches to zero, the mean of it will be very likely to converge into the global maximum point or a near global maximum point of the real posterior. Although the original mean field variational method [23], [7] can only obtain an approximation of the real posterior, the proposed variational MAP algorithm can find the exact optimal or near-optimal MAP estimate. It is an efficient and direct MAP inference algorithm for complex stochastic systems.

3 Kullback-Leibler Divergence between a Gaussian and an Arbitrary p.d.f.

The KL divergence or relative entropy between two probabilistic distribution $g(\mathbf{x})$ and $p(\mathbf{x})$ is defined as

$$KL(g(\mathbf{x})\|p(\mathbf{x})) = \int_{\mathbf{x}} g(\mathbf{x}) \log \frac{g(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}. \tag{1}$$

It is a measurement of the dissimilarity between two distributions. And, it has the property that it is zero if $g(\mathbf{x})$ and $p(\mathbf{x})$ are equal almost everywhere (a.e.) and positive otherwise. But, it is not a real distance since it is not symmetric, i.e., $KL(g(\mathbf{x})\|p(\mathbf{x})) \neq KL(p(\mathbf{x})\|g(\mathbf{x}))$. Generally, minimizing $KL(g(\mathbf{x})\|p(\mathbf{x}))$ with regard to $g(\mathbf{x})$ will favor those $g(\mathbf{x})$ distributions whose probability densities all lie in the regions with high probability under $p(\mathbf{x})$, but without the requirement that all those areas are covered. While minimizing $KL(p(\mathbf{x})\|g(\mathbf{x}))$ with regard to $g(\mathbf{x})$ will favor the settings of $g(\mathbf{x})$ which can cover all the high probability areas in $p(\mathbf{x})$, even if this will result in assigning the high probability area of $g(\mathbf{x})$ to the very low probability area of $p(\mathbf{x})$ [12].

It is also worth noting that the KL divergence in (1) is finite only when $g(\mathbf{x})$ and $p(\mathbf{x})$ have the same support (we set $0 \log \frac{0}{0} = 0$, which is motivated by continuity) [38]. Thus, if $g(\mathbf{x})$ is a Gaussian and $p(\mathbf{x})$ is compactly supported, the $KL(g(\mathbf{x})\|p(\mathbf{x}))$ will be $+\infty$.

Based on the above observations, if we constrain the $g(\mathbf{x})$ distribution to be a Gaussian distribution, we have the following theorem relating the supreme of $p(\mathbf{x})$ and the infimum of $KL(g(\mathbf{x})\|p(\mathbf{x}))$. We must emphasize beforehand that the integrability assumption in (2) is essential; otherwise, the $KL(g(\mathbf{x})\|p(\mathbf{x}))$ could be $+\infty$ no matter how the Gaussian distribution $g(\mathbf{x})$ is translated and scaled.

Theorem 1. *Let $p(\mathbf{x})$, $\mathbf{x}$ is a random vector in $\mathcal{R}^n$, be a bounded, continuous, and everywhere positive p.d.f. with the properties:*

- *There exists a unique $\mathbf{x}^* \in \mathcal{R}^n$ such that $p(\mathbf{x}^*) = \sup_{\mathbf{x}\in\mathcal{R}^n} p(\mathbf{x})$.*
- *$p(\mathbf{x})$ is proper, i.e., $p(\mathbf{x}) \to \mathbf{0}$ as $\mathbf{x} \to \infty$.*
- *The following integrability condition in (2) holds*

$$\left| \int_{\mathbf{x}} exp\left\{ -\frac{\mathbf{x}^T\mathbf{x}}{2} \right\} \log p(\mathbf{x}) d\mathbf{x} \right| < +\infty. \tag{2}$$

Suppose $q(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathcal{I}_n)$ is a Gaussian distribution with zero mean and identity covariance matrix $\mathcal{I}_n$, then denote $q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) \sim \mathcal{N}(\mathbf{x}|\vec{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2\mathcal{I}_n), \mathbf{x} \in \mathcal{R}^n$ as the Gaussian distribution with mean $\vec{\boldsymbol{\mu}}$ and diagonal covariance $\boldsymbol{\sigma}^2\mathcal{I}_n$. Assume $\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}$ is such that $KL(q^{\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}}_{\boldsymbol{\sigma}}(\mathbf{x})\|p(\mathbf{x})) = \inf_{\vec{\boldsymbol{\mu}}} KL(q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x})\|p(\mathbf{x}))$, then

$$\lim_{\boldsymbol{\sigma}\to 0} \vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} = \mathbf{x}^*. \tag{3}$$

Proof. The proof could be found in Appendix 2 based on several Lemmas in Appendix 1. □

Equation (3) in Theorem 1 nicely reveals to us a DA scheme to find the maximum point of $p(\mathbf{x})$, i.e., we can minimize with regard to $\vec{\boldsymbol{\mu}}$ a series of $KL(q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x})\|p(\mathbf{x}))$. This can be achieved by initially setting the $\boldsymbol{\sigma}^2$ to be very large value and decreasing it asymptotically to zero. When the $\boldsymbol{\sigma}^2$ is very large, the optimization of $KL(q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x})\|p(\mathbf{x}))$ is just a convex optimization problem [17]. With the decreasing of the $\boldsymbol{\sigma}^2$, the $KL(q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x})\|p(\mathbf{x}))$ will have more local minima and the optimization is more complex. For a fixed $\boldsymbol{\sigma}^2$, we can run an optimization algorithm to find the minimum of $KL(q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x})\|p(\mathbf{x}))$, then the result will be used as the initial point of the optimization in the next step of annealing. As $\boldsymbol{\sigma}^2$ decreases asymptotically to zero, the whole annealed optimization process will be very likely to converge into the global minimum of the $KL(q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x})\|p(\mathbf{x}))$ and, thus, the global maximum of $p(\mathbf{x})$.

Moreover, in many cases, the $p(\mathbf{x})$ is not directly in hand, so we may not be able to maximize it directly. For example, in the Bayesian inference problem presented in Section 4, where $p(\mathbf{x})$ is corresponding to the posterior distribution which must be inferred from the observations. In Section 5, we show that by using a novel variational inference framework, the problem of optimal MAP estimation can be efficiently solved by minimizing the KL divergence between a variational Gaussian and the real posterior distribution without explicitly recovering the latter.

4 Multivariate Gaussian Constrained Mean Field Variational Analysis

In this section, we present the Gaussian constrained mean field variational analysis, which functions as the optimization method in one annealing step in the variational MAP algorithm. To better illustrate it, we adopt a specific type of graphical model, i.e., the Markov network as shown in Fig. 1. Since different types of graphical models can be transformed to one another [23], adopt a specific type of graphical model will not lose the generality of the proposed algorithm.

In a Markov network, each $\mathbf{z}_i$ represents an observation of the latent random variable $\mathbf{x}_i$. Each undirected link is associated with a potential function $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$, which models the probability of two adjacent nodes being in a certain state pair. And, each directed link represents an observation function $\phi_i(\mathbf{z}_i|\mathbf{x}_i)$ which models the probability of the observation $\mathbf{z}_i$ given $\mathbf{x}_i$. Denotes $\mathbf{X} = \{\mathbf{x}_i, i = 1 \dots \mathcal{L}\}$ as the set of latent random variables and $\mathbf{Z} = \{\mathbf{z}_i, i = 1 \dots \mathcal{L}\}$ as the set of all observations. Then, the joint probability of the Markov network is

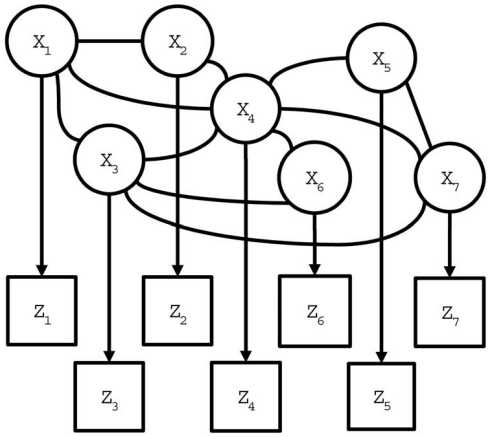

Fig. 1. An example of the Markov network.

$$P(\mathbf{X}, \mathbf{Z}) = \frac{1}{Z} \prod_{\{i,j\}\in\mathcal{E}} \psi_{ij}(\mathbf{x}_i, \mathbf{x}_j) \prod_{i\in\mathcal{V}} \phi_i(\mathbf{z}_i|\mathbf{x}_i), \tag{4}$$

where $\mathcal{E}$ is the set of undirected links, $\mathcal{V}$ is the set of directed links, and Z is a normalization constant. Then, the Bayesian MAP inference in the Markov network is to find

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} P(\mathbf{X}|\mathbf{Z}). \tag{5}$$

We show that by combining the mean field variational method [12], [11], [32], [7], [10] with the DA [16], [17], we can efficiently find the optimal or near optimal MAP estimation of the joint posterior $P(\mathbf{X}|\mathbf{Z})$.

To achieve that, first, we adopt the mean field approximation, i.e.,

$$P(\mathbf{X}|\mathbf{Z}) \approx Q(\mathbf{X}) = \prod_{i=1}^{\mathcal{L}} Q_i(\mathbf{x}_i). \tag{6}$$

Suppose all the random variables share one common dimension N, we further constrain each of the $Q_i(\mathbf{x}_i)$ as a multivariate Gaussian, i.e.,

$$Q_i(\mathbf{x}_i) \sim \mathcal{N}(\mathbf{x}_i|\vec{\boldsymbol{\mu}}_i, \boldsymbol{\Sigma}_i), \tag{7}$$

where $\vec{\boldsymbol{\mu}}_i$ is the N-dimensional mean vector and $\boldsymbol{\Sigma}_i = \boldsymbol{\sigma}^2\mathcal{I}_N$ is the $N \times N$ diagonal covariance matrix. Then, $Q(\mathbf{X})$ is a $N \cdot \mathcal{L}$ dimensional multivariate Gaussian distribution with $N \cdot \mathcal{L} \times N \cdot \mathcal{L}$ diagonal covariance matrix as follows:

$$\begin{aligned} Q(\mathbf{X}) &\sim \mathcal{N}(\mathbf{X}|\vec{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) \\ &= \mathcal{N}\left(\mathbf{X} \middle| \begin{bmatrix} \vec{\boldsymbol{\mu}}_1 \\ \vec{\boldsymbol{\mu}}_2 \\ \cdot \\ \cdot \\ \vec{\boldsymbol{\mu}}_{\mathcal{L}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\sigma}^2\mathcal{I}_N & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma}^2\mathcal{I}_N & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdot & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\sigma}^2\mathcal{I}_N \end{bmatrix} \right). \end{aligned} \tag{8}$$

We can thus construct a cost function, i.e.,

$$J(Q) = \log P(\mathbf{Z}) - KL(Q(\mathbf{X})\|P(\mathbf{X}|\mathbf{Z})) \quad (9)$$

$$= \log P(\mathbf{Z}) - \oint_{\mathbf{X}} \prod_j Q_j(\mathbf{x}_j) \log\Big(\frac{\prod_j Q_j(\mathbf{x}_j)}{P(\mathbf{X}|\mathbf{Z})}\Big) d\mathbf{X} \quad (10)$$

$$= \sum_j H_j(Q_j(\mathbf{x}_j)) + \int_{\mathbf{x}_i} Q_i(\mathbf{x}_i) E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\} d\mathbf{x}_i, \quad (11)$$

where

$$H_j(Q_j(\mathbf{x}_j)) = -\int_{\mathbf{x}_j} Q_j(\mathbf{x}_j) \log Q_j(\mathbf{x}_j) d\mathbf{x}_j \quad (12)$$

is the entropy of the distribution $Q_j(\mathbf{x}_j)$ and

$$E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\} = \oint_{\{\mathbf{x}_j\}\setminus\mathbf{x}_i} \prod_{\{\mathbf{j}\}\setminus\mathbf{i}} Q_j(\mathbf{x}_j) \log P(\mathbf{X},\mathbf{Z}) d\mathbf{X}. \quad (13)$$

Note that (11) holds for any $i = 1 \ldots \mathcal{L}$. It is easy to figure out that maximizing $J(Q)$ is equivalent to minimizing $KL(Q(\mathbf{X})||P(\mathbf{X}|\mathbf{Z}))$ since $P(\mathbf{Z})$ is in fact a constant. We incorporate $\log P(\mathbf{Z})$ in the cost function because we can thus apply the Byesian rule to transform the posterior in (9) to the joint probability in (11). Therefore, as we will demonstrate later, we can obtain more convenience in computation by using the factorized form of the joint probability in (4).

We solve this constrained optimization problem by taking a strategy similar to the gradient projection method [39], [40]. First, we relax the constraint by letting $Q_i(\mathbf{x}_i)$ be any valid probabilistic distributions. Then, we can use the Lagrangian multipliers to reinforce the constraint that $\int_{\mathbf{x}_i} Q_i(\mathbf{x}_i) d\mathbf{x}_i = 1$, i.e.,

$$J^*(Q) = J(Q) + \sum_i \lambda_i \Big(\int_{\mathbf{x}_i} Q_i(\mathbf{x}_i) d\mathbf{x}_i - 1 \Big). \quad (14)$$

Therefore, now we need to minimize the functional $J^*(Q)$. Differentiating it with respect to $Q_i(\mathbf{x}_i)$ and λ_i and setting them to zero, we would obtain the following set of Euler equations, i.e.,

$$\begin{cases} -\log Q_i(\mathbf{x}_i) - 1 + E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\} + \lambda_i = 0 \\ \int_{\mathbf{x}_i} Q_i(\mathbf{x}_i) d\mathbf{x}_i - 1 = 0. \end{cases} \quad (15)$$

To solve this equation set, we easily obtain

$$\begin{cases} Q_i(\mathbf{x}_i) = \exp(\lambda_i - 1) \exp\big(E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\}\big) \\ \lambda_i = 1 - \log\Big(\int_{\mathbf{x}_i} \mathbf{e}^{E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\}}\Big). \end{cases} \quad (16)$$

Thus, we can easily obtain the set of mean field fix-point equations [32], [7], [10] for the updating of $Q_i(\mathbf{x}_i)$ for each $i = 1 \ldots \mathcal{L}$, i.e.,

$$Q_i(\mathbf{x}_i) = \frac{1}{Z_i} \mathbf{e}^{E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\}}, \quad (17)$$

where

$$Z_i = \int_{\mathbf{x}_i} \mathbf{e}^{E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\}} \quad (18)$$

is the normalization constant to assure that $Q_i(\mathbf{x}_i)$ be a valid probability density function. We can iterate this set of fix-point equations in order to find a minimum point of $KL(Q(\mathbf{X})\|P(\mathbf{X}|\mathbf{Z}))$ when $Q(\mathbf{X})$ is a product of $\mathcal{L}$ Gaussian distributions with fixed covariance $\boldsymbol{\sigma}^2\mathcal{I}_n$, i.e.,

$$\vec{\boldsymbol{\mu}}_i = \int_{\mathbf{x}_i} \mathbf{x}_i Q_i(\mathbf{x}_i) d\mathbf{x}_i \quad (19)$$

$$= \frac{1}{Z_i} \int_{\mathbf{x}_i} \mathbf{x}_i \mathbf{e}^{E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\}} d\mathbf{x}_i. \quad (20)$$

In fact, it is easy to figure out that (20) will minimize the $KL(Q_i(\mathbf{x}_i)||\mathcal{N}(\mathbf{x}_i|\vec{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2\mathcal{I}_n))$ with regard to $\mathcal{N}(\mathbf{x}_i|\vec{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2\mathcal{I}_n)$, where $Q_i(\mathbf{x}_i)$ is the unconstrained variational p.d.f. from (17) and $\mathcal{N}(\mathbf{x}_i|\vec{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2\mathcal{I}_n)$ is a Gaussian distribution with fixed covariance $\boldsymbol{\sigma}^2\mathcal{I}_n$. In this sense, (20) represents a projection of any p.d.f. $Q_i(\mathbf{x}_i)$ to the functional space spanned by all the Gaussian distributions with the fixed covariance $\boldsymbol{\sigma}^2\mathcal{I}_n$. Then, the projected Gaussian distribution will be used for the next mean field iteration. This process will continue until the mean field iterations reach the fix-point. It exactly follows the same strategy of the gradient projection method [39], [40].

Embedding (4) and (7) into (20), we obtain the set of factorized fix-point equations, i.e.,

$$\vec{\boldsymbol{\mu}}_i = \frac{1}{Z_i'} \int_{\mathbf{x}_i} \mathbf{x}_i \phi_i(\mathbf{z}_i|\mathbf{x}_i) \mathbf{e}^{\sum_{j\in\mathcal{N}(i)} \int_{\mathbf{x}_j} \mathcal{N}(\mathbf{x}_j|\vec{\boldsymbol{\mu}}_j, \boldsymbol{\sigma}^2\mathcal{I}_N) \log \psi_{ij}(\mathbf{x}_i,\mathbf{x}_j) d\mathbf{x}_j} d\mathbf{x}_i, \quad (21)$$

where

$$Z_i' = \int_{\mathbf{x}_i} \phi_i(\mathbf{z}_i|\mathbf{x}_i) \mathbf{e}^{\sum_{j\in\mathcal{N}(i)} \int_{\mathbf{x}_j} \mathcal{N}(\mathbf{x}_j|\vec{\boldsymbol{\mu}}_j, \boldsymbol{\sigma}^2\mathcal{I}_N) \log \psi_{ij}(\mathbf{x}_i,\mathbf{x}_j) d\mathbf{x}_j} d\mathbf{x}_i \quad (22)$$

is again a normalization constant and $\mathcal{N}(i)$ indicates the set of neighboring nodes of $\mathbf{x}_i$. Then, we iteratively assign $Q_i(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i|\vec{\boldsymbol{\mu}}_i, \boldsymbol{\sigma}^2\mathcal{I}_N)$, where $\vec{\boldsymbol{\mu}}_i$ is calculated according to (21). Please note that the covariance of the variational Gaussian distribution is kept fixed during the fix-point iteration and projection process.

For a constant $\boldsymbol{\Sigma} = \boldsymbol{\sigma}^2\mathcal{I}_{N\mathcal{L}}$, (21) is the mean field fix-point equation to update $\vec{\boldsymbol{\mu}}_i$. We can iterate this set of fix-point equations and $\vec{\boldsymbol{\mu}}_i$ will converge to a minimum point of $KL(Q(\mathbf{X})||P(\mathbf{X}|\mathbf{Z}))$. This set of fix-point equations is efficient since the updating of each $\vec{\boldsymbol{\mu}}_i$ only involves the local computation in the neighborhood of $\mathbf{x}_i$ in the graphical model.

However, another issue of interest is that to solve the constrained maximization of $J(Q)$, we may directly take the derivative of $J(Q)$ with regard to the mean $\vec{\boldsymbol{\mu}}_i$ of each of the Gaussian $Q_i(\mathbf{x}_i)$ and set them to zero. By interchanging the derivative and integral in (11), we can then obtain the following equations

$$\vec{\boldsymbol{\mu}}_i = \frac{\int_{\mathbf{x}_i} \mathbf{x}_i Q_i(\mathbf{x}_i) E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\} d\mathbf{x}_i}{\int_{\mathbf{x}_i} Q_i(\mathbf{x}_i) E_Q\{\log P(\mathbf{X},\mathbf{Z})|\mathbf{x}_i\} d\mathbf{x}_i}. \quad (23)$$

Again, by embedding (4) into (23), we obtain the factorized version of (23), i.e.,

$$\vec{\boldsymbol{\mu}}_i = \frac{1}{Z_i''} \int_{\mathbf{X}} \mathbf{x}_i \prod_{j\in\mathcal{V}} Q_j(\mathbf{x}_j) \left(\sum_{(k,l)\in\mathcal{E}} \log \psi_{kl}(\mathbf{x}_k,\mathbf{x}_l) + \sum_{m\in\mathcal{V}} \log \phi_m(\mathbf{x}_m) \right) d\mathbf{X}, \quad (24)$$

Variational Maximum a Posteriori Algorithm

1) `Initialization:` Annealing control parameter $m = 0$; T_{max} be very large where the annealing starts and T_{min} be very small near zero where the annealing stops; let $\mathcal{I}_N$ be the $N \times N$ identity matrix; denote $\vec{\mu}_{i,0}$ as the set of randomly generated mean vector for initialization.
2) `Annealing:` $m = m+1$, $T = \frac{T_{max}}{m}$, then $\mathbf{\Sigma}_i = T \times \mathcal{I}_N$; $\vec{\mu}_{i,m} = \vec{\mu}_{i,m-1}$, $i = 1 \dots \mathcal{L}$; If $T > T_{min}$, goto Step 3, else goto Step 4.
3) `Mean field iteration:` Update $\vec{\mu}_{i,m}$ based on the current value of $\vec{\mu}_{i,m}$ and the fixed $\mathbf{\Sigma}_i$ according to Equation 21. Iterate this step until convergence. Then jump back to Step 2.
4) `Result:` $\vec{\mu}_i^* = \vec{\mu}_{i,m}$, $i = 1 \dots \mathcal{L}$, are the MAP estimation of $P(\mathbf{X}|\mathbf{Z})$.

Fig. 2. Variational MAP algorithm.

where

$$Z_i'' = \int_{\mathbf{X}} \prod_{j \in \mathcal{V}} Q_j(\mathbf{x}_j) \left(\sum_{(k,l) \in \mathcal{E}} \log \psi_{kl}(\mathbf{x}_k, \mathbf{x}_l) + \sum_{m \in \mathcal{V}} \log \phi_m(\mathbf{x}_m) \right) d\mathbf{X} \tag{25}$$

is a normalization constant.

While it seems that (24) be a more direct solution, our experiments show that even in a relative simple synthetic problem as that in Section 7.2, the iteration of (24) failed to converge. Two reasons might explain why this happens: 1) the deduction of (23) involves an interchange between derivative and integral, which may not be justified and 2) The iteration of (24) is not numerically stable, i.e., it might be easily got trapped in some saddle points. Another reason that we adopt (19) is that the updating of $\vec{\boldsymbol{\mu}}_i$ only involves the local computation in the neighborhood of the node $\mathbf{x}_i$, while (24) does not have such kind of nice local property. Therefore, (19) is more justified as well as more computational efficient than (24).

5 Variational MAP by Deterministic Annealing

Based on Theorem 1 in Section 3 and the multivariate Gaussian constrained mean field variational analysis in Section 4, we show that we can nicely adopt a DA scheme to efficiently find the optimal MAP estimate without explicitly recovering the $P(\mathbf{X}|\mathbf{Z})$.

We first relax the problem of estimating the global maximum of $P(\mathbf{X}|\mathbf{Z})$, i.e., we can instead minimize $KL(Q(\mathbf{X}) \| P(\mathbf{X}|\mathbf{Z}))$, where $Q(\mathbf{X})$ is constrained to be a multivariate Gaussian with a fixed diagonal covariance $\mathbf{\Sigma} = \boldsymbol{\sigma}^2 \mathcal{I}_{N\mathcal{L}}$ as in (8). We can then apply the DA scheme revealed by Theorem 1. This is achieved by regarding the $\boldsymbol{\sigma}^2$ as the temperature T for annealing. We can set it to be very large at the start. The minimization of the $KL(Q(\mathbf{X}) \| P(\mathbf{X}|\mathbf{Z}))$ in this start setting is usually a trivial convex optimization problem [17]. Then, the multivariate Gaussian constrained mean field iteration in (21) can usually find the only minimum point under this setting. Using this result as an initialization, we decrease $\boldsymbol{\sigma}^2$ to be smaller toward zero and run the mean field iteration in (21) again. We can repeat the process until the $\boldsymbol{\sigma}^2$ decreasing to near zero. Then, upon convergence, the whole annealing process will be very likely to obtain the global minimum of the $\lim_{\boldsymbol{\sigma} \to 0} KL(Q(\mathbf{X}) \| P(\mathbf{X}|\mathbf{Z}))$ and, thus, the global maximum of $P(\mathbf{X}|\mathbf{Z})$. Therefore, we only need to control one parameter $T = \boldsymbol{\sigma}^2$ for the annealing process. Generally, we propose the variational MAP algorithm as shown in Fig. 2.

Nevertheless, the annealing scheme, i.e., the decreasing rate of T, needs to be carefully designed to have a good optimization result. Unfortunately, it seems that a theoretic analysis of the annealing rate is very difficult. In the proposed algorithm, we let the annealing control parameter T decrease hyperbolically with the annealing number K. In our experiments, such an annealing scheme always obtains satisfactory results. Please note that although the mean field variational analysis can only obtain an approximate posterior, the proposed algorithm is very likely to obtain the exact optimal MAP estimate.

6 Monte Carlo Simulation of the Variational MAP

In a real valued graphical model such as that in Fig. 1, if all the observation functions $\phi_i(\mathbf{z}_i|\mathbf{x}_i)$ and all the potential functions $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ are Gaussian, then we may obtain a closed form analytical solution of the fix-point equations in (21). However, either the $\phi_i(\mathbf{z}_i|\mathbf{x}_i)$ or the $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ could be complex non-Gaussian distributions, e.g., the image observation function in the CONDENSATION contour tracker [29], [30], [31] is the interference of a Gaussian random process and a Poisson random process due to the background clutter. This makes it very difficult to obtain analytical solutions for the fix-point equations in (21), e.g., it would be very difficult to evaluate the normalization constant Z_i' in (22) since it involves multiple integrals of complex distributions.

Nevertheless, under the non-Gaussian case, we can seek the help of Monte Carlo simulation to approximately evaluate (21). According to the strong law of large numbers, as the number of i.i.d. samples from a distribution approaches to infinity, any order of the sample quadrature will converge to the same order of distribution statistics with probability one. Thus, to evaluate (21), first, we can generate $\mathcal{L}$ sets of samples to approximate each of the $Q_i(\mathbf{x}_i)$, i.e.,

Variational MAP Monte Carlo

1) `Initialization:` Set the annealing control parameter $m = 0$; let T_{max} be very large where the annealing starts and T_{min} be very small near zero where the annealing stops; let $\mathcal{I}_N$ be the $N \times N$ identity matrix; denote $\vec{\mu}_{i,0}$ as the set of randomly generated mean vector for initialization.
2) `Annealing:` $m = m+1$, $T = \frac{T_{max}}{m}$, then $\mathbf{\Sigma}_i = T \times \mathcal{I}_N$; $\vec{\mu}_{i,m} = \vec{\mu}_{i,m-1}$, $i = 1 \ldots \mathcal{L}$. If $T > T_{min}$, goto Step 3, else goto Step 4.
3) `Mean field iteration:` Sample $\{\mathbf{s}_{i,k}\}_{k=1}^{K}$ from $Q_i(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i|\vec{\mu}_{i,m}, \mathbf{\Sigma}_i)$ for $i = 1 \ldots \mathcal{L}$, and calculate the updated $\vec{\mu}_{i,m}$ based on these sets of samples according to Equation 27. Iterate this step until convergence. Then jump back to Step 2.
4) `Result:` $\vec{\mu}_i^* = \vec{\mu}_{i,m}$, $i = 1 \ldots \mathcal{L}$, are the MAP estimation of $P(\mathbf{X}|\mathbf{Z})$.

Fig. 3. Monte Carlo implementation of the variational MAP algorithm.

$$Q_i(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i|\vec{\boldsymbol{\mu}}_i, \boldsymbol{\sigma}^2\mathcal{I}_N) \sim \{\mathbf{s}_{i,k}\}_{k=1}^{K}, i = 1 \ldots \mathcal{L}, \quad (26)$$

where K is the number of samples used for simulation. Then, these $\mathcal{L}$ sets of samples can be used for evaluating (21) approximately, i.e.,

$$\vec{\boldsymbol{\mu}}_i = \frac{1}{Z_i'''} \sum_{k=1}^{K} \mathbf{s}_{i,k} \phi_i(\mathbf{z}_i|\mathbf{s}_{i,k}) \exp\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{K} \sum_{l=1}^{K} \log \psi_{ij}(\mathbf{s}_{i,k}, \mathbf{s}_{j,l}) \right), \quad (27)$$

where

$$Z_i''' = \sum_{k=1}^{K} \phi_i(\mathbf{z}_i|\mathbf{s}_{i,k}) \exp\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{K} \sum_{l=1}^{K} \log \psi_{ij}(\mathbf{s}_{i,k}, \mathbf{s}_{j,l}) \right) \quad (28)$$

is the normalization constant. Therefore, we propose the Monte Carlo implementation of the variational MAP algorithm in Fig. 3.

In fact, the use of Monte Carlo simulation in the variational MAP algorithm has other advantages in some computer vision applications. For example, in visual tracking, since the detection of the target is, in general, very difficult, it would be hard to obtain the image observation $\mathbf{z}_i$ and, thus, it is hard to evaluate the observation likelihood $\phi_i(\mathbf{z}_i|\mathbf{x}_i)$. Whereas in a sample-based Monte Carlo algorithm, the observation likelihood $\phi(\mathbf{z}_i|\mathbf{x}_i)$ can be evaluated in a top-down approach, i.e., for each sample $\mathbf{s}_{i,k}$, we can easily match the model represented by the sample with the image data or image features corresponding to the sample, just as in the CONDENSATION contour tracker [29], [30], [31].

7 Experiments

In this section, we present extensive experimental results of both synthetic problems and real applications, which demonstrate the effectiveness and efficiency of the proposed variational MAP algorithm.

7.1 Evolution of the Topology of the KL Divergence during Annealing

In this experiment, we use an illustrative example to present the topology of the KL divergence between a Gaussian distribution and a multimodal Gaussian mixture during the process of annealing. As shown in Fig. 4, it does evolve as we expected from Theorem 1.

The real distribution $p(\mathbf{x}) = 0.4 \cdot \mathcal{N}(\mathbf{x}| - 5, 0.2) + 0.1 \cdot \mathcal{N}(\mathbf{x}| - 2, 0.2) + 0.25 \cdot \mathcal{N}(\mathbf{x}|3, 0.2) + 0.25 \cdot \mathcal{N}(\mathbf{x}|5, 0.2)$ is a Gaussian mixture of four kernels. The $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ is a Gaussian distribution. The annealing parameter is $T = \boldsymbol{\sigma}^2$. We can observe in Fig. 4a that when T is large, i.e., $T = 16.0$, the $KL(q(\mathbf{x})\|p(\mathbf{x}))$ is really a convex function with regards to $\boldsymbol{\mu}$. Then, with the decreasing of T, the $KL(q(\mathbf{x})\|p(\mathbf{x}))$ will have more local minima, i.e., when $T = 6.0$ or $T = 2.0$, the $KL(q(\mathbf{x})\|p(\mathbf{x}))$ has two local minima as shown in Figs. 4b and 4c. As the T decreases asymptotically to near zero, i.e., $T = 0.2$, the $KL(q(\mathbf{x})\|p(\mathbf{x}))$ has four local minima at $\boldsymbol{\mu} = -5.0, -2.0, 3.0, 5.0$, respectively. Each of them corresponds to one of the four local maxima of $p(\mathbf{x})$ at $\boldsymbol{\mu} = -5.0, -2.0, 3.0, 5.0$. Also, the global minimum of the $KL(q(\mathbf{x})\|p(\mathbf{x}))$ is at $\boldsymbol{\mu} = -5.0$, which exactly corresponds to the global maximum of $p(\mathbf{x})$ at $\mathbf{x} = -5.0$, as shown in Fig. 4d.

For comparison, we also present the plot of $p(\boldsymbol{\mu})$ in Fig. 4e and $-\log p(\boldsymbol{\mu})$, in Fig. 4f. Compare Fig. 4d with Fig. 4f, we empirically demonstrate that as a function of $\boldsymbol{\mu}$, the topology of $KL(q(\mathbf{x})\|p(\mathbf{x}))$ does converge to the topology $-\log p(\boldsymbol{\mu})$, as $\boldsymbol{\sigma}^2$ approaches to zero. This result is what we expect from the Lemma 2 of Theorem 1 in the appendix.

7.2 Variational MAP inference in an Illustrative Synthetic Problem

To investigate the convergence of the proposed variational MAP algorithm, we perform it on an illustrative synthetic problem, which is modeled as a two-nodes Markov network in Fig. 5. In this synthetic problem, both $\mathbf{x}_1$ and $\mathbf{x}_2$ are one-dimensional random variables. The potential function between these two random variables is modeled as a Gaussian distribution, i.e.,

$$\psi_{12}(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2 - \mathbf{x}_1|6.0, 0.3). \quad (29)$$

The observation function $\phi_i(\mathbf{z}_i|\mathbf{x}_i)$, $i = 1, 2$ are modeled as two Gaussian mixtures, respectively, i.e.,

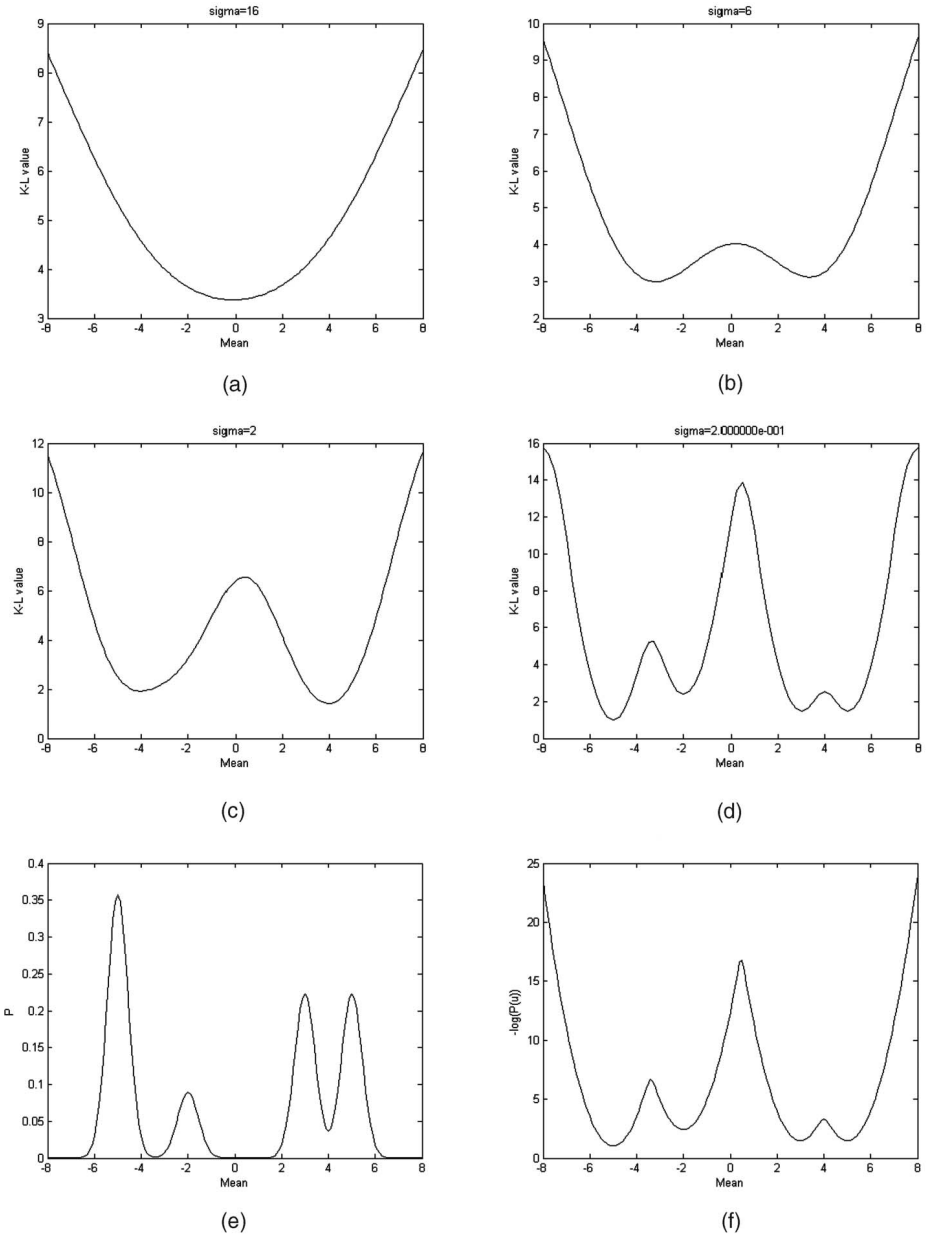

Fig. 4. (a) $T = \boldsymbol{\sigma}^2 = 16.0$. (b) $T = \boldsymbol{\sigma}^2 = 6.0$. (c) $T = \boldsymbol{\sigma}^2 = 2.0$. (d) $T = \boldsymbol{\sigma}^2 = 0.2$. (e) $p(\boldsymbol{\mu})$. (f) $-log(p(\boldsymbol{\mu}))$. Evolution of the $KL(q(\mathbf{x})\|p(\mathbf{x}))$ with regard to $\boldsymbol{\mu}$ during annealing, where $p(\mathbf{x}) = 0.4 \cdot \mathcal{N}(\mathbf{x}| -5, 0.2) + 0.1 \cdot \mathcal{N}(\mathbf{x}| -2, 0.2) + 0.25 \cdot \mathcal{N}(\mathbf{x}|3, 0.2) + 0.25 \cdot \mathcal{N}(\mathbf{x}|5, 0.2)$, $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ and $T = \boldsymbol{\sigma}^2$: (a) The KL topology when $T = 16.0$, (b) the KL topology when $T = 6.0$, (c) the KL topology when $T = 2.0$, (d) the KL topology when $T = 0.2$, (e) the plot of the Gaussian mixture $p(\boldsymbol{\mu})$, and (f) the plot of the $-\log p(\boldsymbol{\mu})$.

$$\begin{aligned}\phi_1(\mathbf{z}_1|\mathbf{x}_1) &= 0.5\mathcal{N}(\mathbf{z}_1 - \mathbf{x}_1| -3.0, 0.3) + 0.4\mathcal{N}(\mathbf{z}_1 - \mathbf{x}_1|0, 0.2) \\ &\quad + 0.1\mathcal{N}(\mathbf{z}_1 - \mathbf{x}_1|4, 0.4) \qquad (30)\\ \phi_2(\mathbf{z}_2|\mathbf{x}_2) &= 0.3\mathcal{N}(\mathbf{z}_2 - \mathbf{x}_2| -5.0, 0.2) \\ &\quad + 0.1\mathcal{N}(\mathbf{z}_2 - \mathbf{x}_2| -2.0, 0.3) \\ &\quad + 0.4\mathcal{N}(\mathbf{z}_2 - \mathbf{x}_2|3.0, 0.2) \\ &\quad + 0.2\mathcal{N}(\mathbf{z}_2 - \mathbf{x}_2|5.0, 0.1). \qquad (31)\end{aligned}$$

Then, we randomly choose the observations $\mathbf{z}_1$ and $\mathbf{z}_2$ and perform the proposed variational MAP algorithm on it, we show the Bayesian MAP inference results in Fig. 6.

From Fig. 6a, we can observe the convergence of the proposed variational MAP algorithm in this illustrative synthetic problem when $\{\mathbf{z}_1, \mathbf{z}_2\} = \{10.0, 16.0\}$. We randomly choose the initialization of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and run the algorithm many times, every time we obtain the same convergence curve, i.e., the converged result after the first step of annealing will always be the "$*$" shown in Fig. 6a at $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{9.6011, 17.6728\}$. This is what we expected since when T is very large, the $KL(\cdot)$ is a convex function and, thus, the optimization in this case will surely converge into the only minimum point, e.g., $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{9.6011, 17.6728\}$ in this case.

We can also observe that the proposed algorithm does converge to the global maximum of the posterior distribution, i.e., our algorithm converges at $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{12.6824, 18.3793\}$ which is shown as the "$\triangle$" in Fig. 6a and the numerically calculated MAP estimate is at around $\{\mathbf{x}_1, \mathbf{x}_2\} = \{12.70, 18.40\}$. Considering the possible error of the numerically calculated MAP estimate, we conclude that our algorithm does recover the global maximum of the

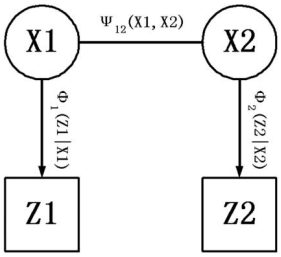

Fig. 5. Two nodes Markov network for the illustrative synthetic problem, where $\psi_{12}(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{N}(x_2 - \mathbf{x}_1|6.0, 0.3)$, $\phi_1(\mathbf{z}_1|\mathbf{x}_1) = 0.5\mathcal{N}(\mathbf{z}_1|\mathbf{x}_1 - 3.0, 0.3) + 0.4\mathcal{N}(\mathbf{z}_1|\mathbf{x}_1, 0.2) + 0.1\mathcal{N}(\mathbf{z}_1|\mathbf{x}_1 + 4, 0.4)$ and $\phi_2(\mathbf{z}_2|\mathbf{x}_2) = 0.3\mathcal{N}(\mathbf{z}_2|\mathbf{x}_2 - 5.0, 0.2) + 0.1\mathcal{N}(\mathbf{z}_2|\mathbf{x}_2 - 2.0, 0.3) + 0.4\mathcal{N}(\mathbf{z}_2|\mathbf{x}_2 + 3.0, 0.2) + 0.2\mathcal{N}(\mathbf{z}_2|\mathbf{x}_2 + 5.0, 0.1)$.

posterior distribution $P(\mathbf{x}_1, \mathbf{x}_2|\mathbf{z}_1 = 10.0, \mathbf{z}_2 = 16.0)$. For comparison and visualization, we also present the topology of the posterior distribution $P(\mathbf{x}_1, \mathbf{x}_2|\mathbf{z}_1 = 10.0, \mathbf{z}_2 = 16.0)$ in Fig. 6b.

Although in theory, we cannot guarantee the algorithm to obtain the global optimal MAP estimation, extensive running of the experiments on the synthetic problem shows that the proposed variational MAP algorithm does always converge to the global maximum of the posteriori distribution. We present two other experimental results from Fig. 6c to Fig. 6f. Again, both the convergence curve and the topology of the posterior distribution are presented.

Some of the details of the experiments are described as follows: First, the T_{max} is set to 200 and T_{min} is set to 0.01, where the annealing starts and ends respectively. Second, in each step of the annealing, we iterate (21) until convergence, i.e., we stop the updating of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ if the difference between the updated value and the previous value is below the prespecified threshold of 0.01.

Another concern would be about the convergence rate of the proposed variational MAP algorithm. Although a theoretical analysis of the convergence rate is very difficult, on the synthetic two-node problem, we generally observe that the first step of annealing takes the most number of iterations which ranges from 10 to 15 to converge, then in the following steps of annealing, it only takes one to two steps for the mean field iteration to converge. Therefore, empirically we achieve fast convergence of the proposed variational MAP algorithm. By the way, how to design the annealing scheme to achieve better result is also of interest just as we have mentioned in Section 5. However, a theoretic study of this problem seems to be a tremendous work. In all the experiments, we use the hyperbolical decreasing annealing scheme, i.e., $T = \frac{T_{max}}{K}$, it does achieve satisfactory results.

In fact, instead of manually setting a T_{min} for stopping the annealing, we can develop more rigorous criterion for the convergence of the annealing from the change of the $KL(\cdot)$. To make it clear, we plot the change of the $KL(\cdot)$ during the annealing of the experiment reported in Figs. 6a and 6b, as shown in Fig. 7. From Fig. 7, we observe dramatic decrease of the $KL(\cdot)$ value in the approximately first 2,000 round of annealing. Then, the $KL(\cdot)$ will increase very slowly with the decreasing of T. The hexagons in the plot represents the $KL(\cdot)$ value after each $1,000$ round of annealing. Thus, there is one and only one global minimum $KL(\cdot)$ value during annealing in all the annealing steps. By checking the simulation results, we find that after the annealing which achieves the global minimum $KL(\cdot)$ value, the proposed variational MAP algorithm has already converged to the global maximum of the real posterior, e.g., in the experiments shown in Fig. 7, when the algorithm achieves the global minimum $KL(\cdot)$ value during the annealing, it has converged to the global MAP of the real posteriori distribution at $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{12.6824, 18.3793\}$, which corresponds to the 1,303 round of annealing with $T = \frac{T_{max}}{1,303} = \frac{200}{1,303} = 0.1535$ and the total number of the mean field iteration is 1,420. Actually, in the experiment, the running of the mean field iteration with annealing temperature after $T = 0.1535$ will not change $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ any more, it will just increase the $KL(\cdot)$ value a little bit since the Gaussian variational distribution tends to be more peaky.

Although we only show one plot of the change of the $KL(\cdot)$ value during annealing, all the experiments we have run showed the same pattern of the changes. Therefore, we conclude that we can stop the annealing when we find that after one step of annealing, the resulted $KL(\cdot)$ value is not less than the $KL(\cdot)$ value after the previous step of annealing. This also finds the optimal T_{min} which will result in the most efficient running of the algorithm. However, evaluating the $KL(\cdot)$ value may involve tremendous computation by itself. Therefore, we still tend to manually set the T_{max} and T_{min} to avoid the overhead introduced by the evaluation of the $KL(\cdot)$ value.

Under the same experimental setting, we also run the iteration of (24) on the same problem. Our observation is that the annealed iteration process does not converge at all. We show the experimental results when $\mathbf{z}_1 = 10.0, \mathbf{z}_2 = 16.0$ in Fig. 8. Fig. 8a shows the curve of the annealed iteration of (24), it failed to converge. Checking the value of the KL divergence during the iteration process, we find that it is increasing instead of decreasing with the iteration. We show the curve of the KL divergence in Fig. 8c, while Fig. 8d presents the same curve in the first 100 iteration.

7.3 Variational MAP for Tracking Articulated Human Body

In this experiment, we implement the Monte Carlo simulation of the variational MAP algorithm for tracking an articulated human body. We adopt the same Markov network to represent the articulated human body just as that in [7], where each body part is represented as a quad shape and the motion of each of them is represented as a probabilistic random variable in the six-dimensional affine space. We refer the interested readers to [7] for the detailed description of the potential function $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ and the observation function $\phi_i(\mathbf{z}_i|\mathbf{x}_i)$ of the Markov network.

Then, the Monte Carlo version of the variational MAP algorithm is performed sequentially to recover the motion of the articulated human body from the video. Some of the sample result images are shown in Fig. 9. The proposed variational MAP algorithm recovers the articulated full-body motion very well across the video sequence,[1] which has 767 frames. This is actually the annealed version of the MFMC algorithm proposed in [7], [10].

1. Online demo at http://www.ece.northwestern.edu/~ganghua/PAMI/VMapArticulate.avi.

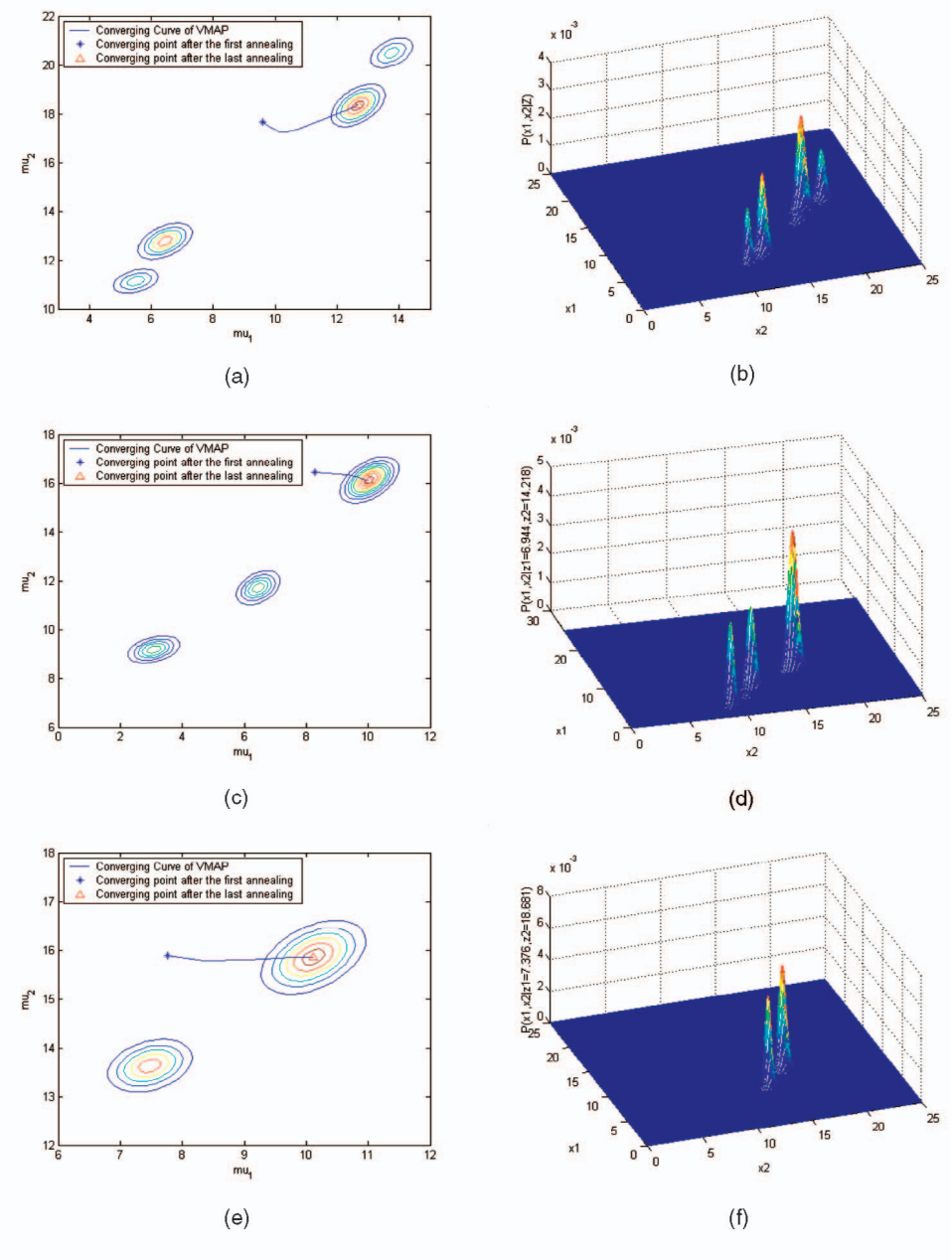

Fig. 6. Convergence of the variational MAP algorithm in the 2D illustrative synthesized problem. The curve in each graph represents the process of convergence. The "$*$" represents the converged result after the first step of annealing, no matter what is the initialization. The "$\triangle$" represents the converged result after the last step of annealing: (a) $\mathbf{z}_1 = 10.0, \mathbf{z}_2 = 16.0$, $* = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{9.6011, 17.6728\}$, $\triangle = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{12.6824, 18.3793\}$. The numerically global maximum is around $\{\mathbf{x}_1, \mathbf{x}_2\} = \{12.70, 18.40\}$, (b) $P(\mathbf{x}_1, \mathbf{x}_2|\mathbf{z}_1 = 10.0, \mathbf{z}_2 = 16.0)$, (c) $\mathbf{z}_1 = 6.944, \mathbf{z}_2 = 14.218$, $* = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{8.2939, 16.4515\}, \triangle = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{10.0353, 16.1268\}$. The numerically global maximum is around $\{\mathbf{x}_1, \mathbf{x}_2\} = \{10.00, 16.10\}$, (d) $P(\mathbf{x}_1, \mathbf{x}_2|\boldsymbol{\mu}_1 = 6.944, \mathbf{z}_2 = 14.218)$. (e) $\mathbf{z}_1 = 7.3762, \mathbf{z}_2 = 18.6813$, $* = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{7.7587, 15.8893\}$, $\triangle = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\} = \{10.1100, 15.8521\}$. The numerically global maximum is around $\{\mathbf{x}_1, \mathbf{x}_2\} = \{10.10, 15.80\}$, (f) $P(\mathbf{x}_1, \mathbf{x}_2|\mathbf{z}_1 = 7.3762, \mathbf{z}_2 = 18.6813)$.

For comparison, we also have implemented the MFMC algorithm [7] and the multiple independent tracker which has been used as a comparison of the MFMC algorithm in [7]. Our experimental results reveal that the MFMC algorithm can track the articulated motion well until the 368th frame and it loses track after that. Sample result images are shown in Fig. 10. For clear visualization, the mean estimate of each of the quadrangle body shapes is overlayed on the images as the tracking results. The reason for the tracking failure of the MFMC algorithm is that the heavy multimodality of the motion posterior causes the mean estimate to be significantly deviated from the MAP estimate of the motion. Thus, it could hardly indicate the true motion, e.g., as we can observe in frame #370. Also, just as reported in [7], the multiple independent tracker loses track from the start.

When comparing the variational MAP algorithm with the MFMC algorithm, we set all the parameters of the potential functions $\psi_{ij}(\mathbf{x}_i, \mathbf{x}_j)$ and the observation functions $\phi_i(\mathbf{z}_i|\mathbf{x}_i)$ to be the same for both algorithms. Because of the annealing process, the proposed variational MAP algorithm needs more mean field iterations than the MFMC algorithm. Our experiments show that only the first step of annealing needs more iterations, in the following steps of annealing, it generally

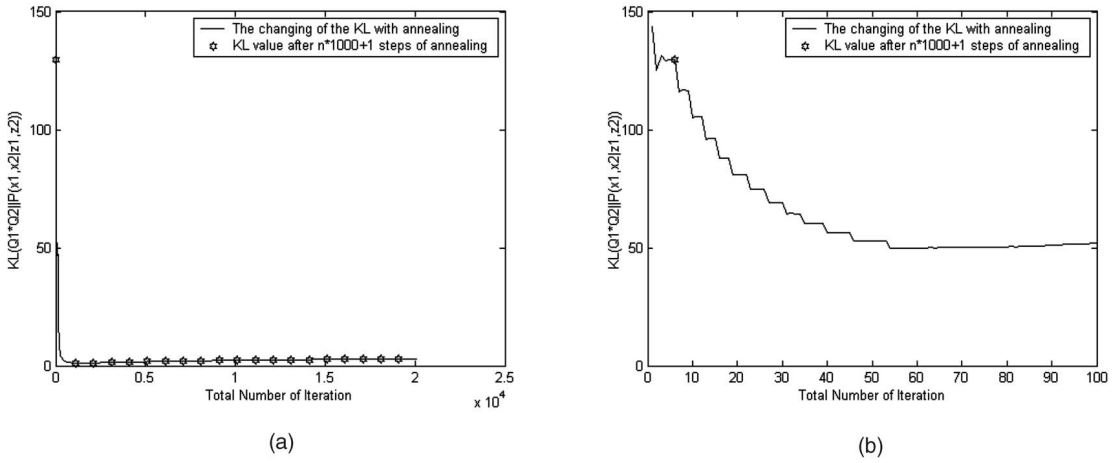

Fig. 7. The change of the $KL(Q_1(\mathbf{x}_1)Q_2(\mathbf{x}_2)\|P(\mathbf{x}_1,\mathbf{x}_2|\mathbf{z}_1=10.0,\mathbf{z}_2=16.0)))$ during annealing. It is dramatically decreased in the first 1,303 round of annealing and then increase very slowly in the following annealing process. The proposed variational MAP algorithm actually has converged to the global maximum of the real posterior distribution $P(\mathbf{x}_1,\mathbf{x}_2|\mathbf{z}_1=10.0,\mathbf{z}_2=16.0))$ at $\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\}=\{12.6824,18.3793\}$ after the 1,303 round of annealing at $T=0.1535$. The total number of the Gaussian constrained mean field iteration is 1,420 up to the end of the 1,303 annealing. (a) All the iterations. (b) First 100 iterations.

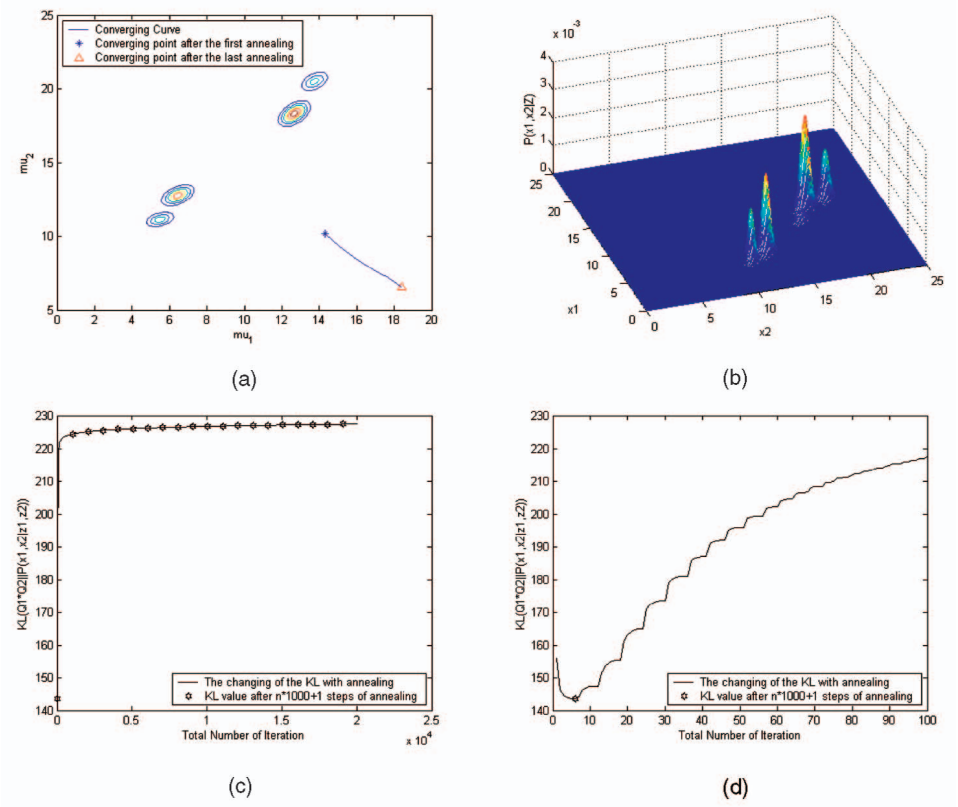

Fig. 8. Annealed iteration of (24) in the 2D illustrative synthesized problem. The "$*$" represents the result after the first step of annealing. And the "$\triangle$" represents the result after the last step of annealing. The iteration failed to converge: (a) $\mathbf{z}_1=10.0,\mathbf{z}_2=16.0$. $*=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\}=\{14.330,10.148\}$. $\triangle=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\}=\{18.391,6.5445\}$. The numerically global optimal is around $\{\mathbf{x}_1,\mathbf{x}_2\}=\{12.70,18.40\}$, (b) $P(\mathbf{x}_1,\mathbf{x}_2|\mathbf{z}_1=10.0,\mathbf{z}_2=16.0)$, and (c) the KL value change in all the iterations. (d) The KL value change in the first 100 iterations.

needs less than half of the mean field iterations of the MFMC algorithm. So, the variational MAP algorithm only increases the computation linearly in comparison with the MFMC algorithm. Therefore, based on the analysis of the complexity of the MFMC algorithm [7], [10], the variational MAP algorithm also achieves linear complexity with respect to the number of body parts in tracking the articulated human body.

All the algorithms are implemented using C++, no code optimization is performed. They are running in a 2.5 GHz PC under Windows XP. We design six annealing steps and in the first step of the annealing, we iterate the mean field fix-point equations for six times and in the following annealing steps, we run the mean field fix-point equations for three times. The algorithm can thus run at the speed of 0.2 frames

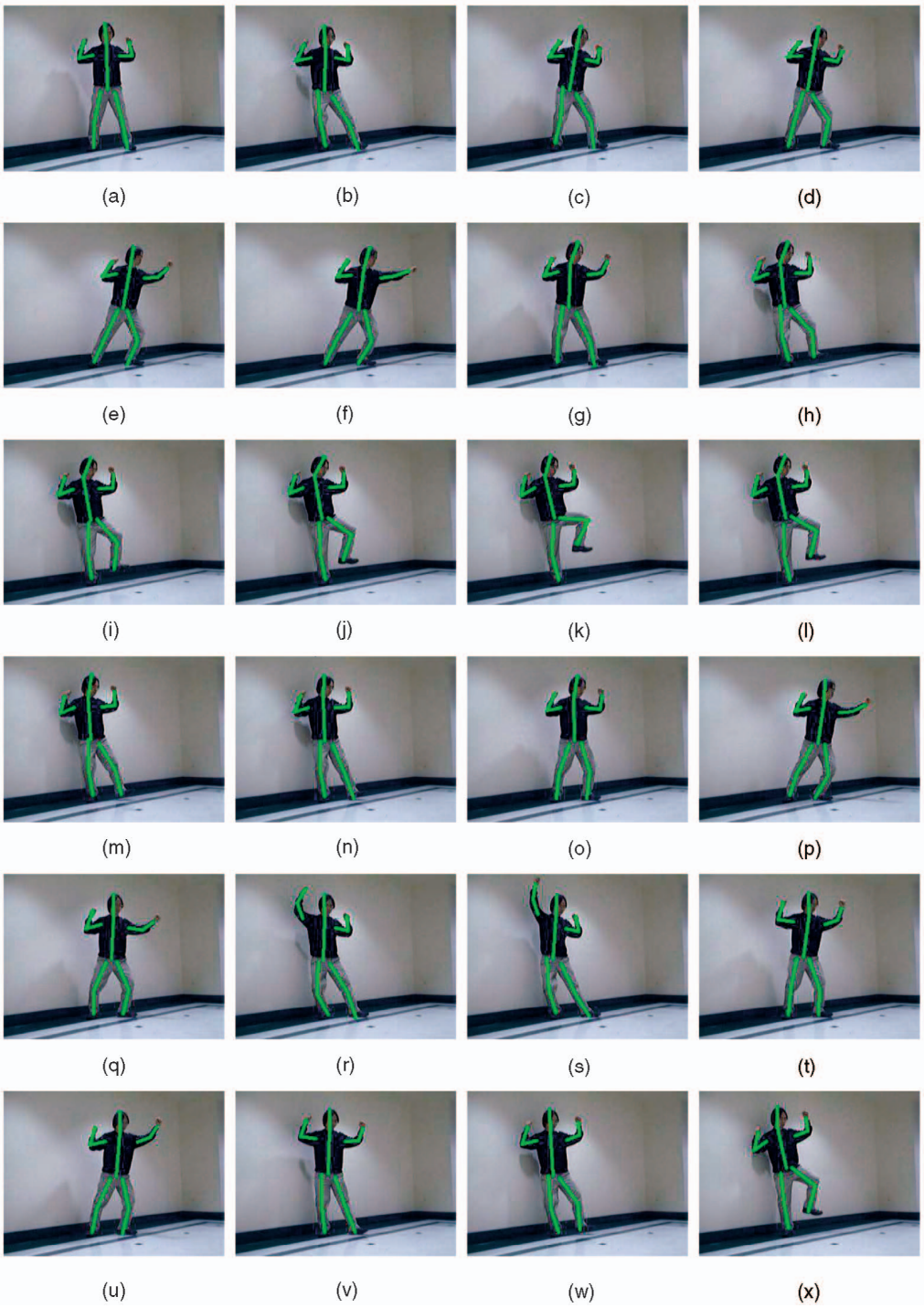

Fig. 9. Variational MAP for tracking articulated human body, the video sequence has 767 frames and it can robustly recover the full human body motion across the whole sequence. We overlay the middle line of each quad shape representing each body parts in the result images. While the MFMC algorithm loses track after 368 frames and the multiple independent tracker loses track from the start. (a) #38, (b) #66, (c) #103, (d) #114, (e) #138, (f) #168, (g) #228, (h) #280, (i) #288, (j) #299, (k) #333, (l) #349, (m) #378, (n) #388, (o) #424, (p) #456, (q) #504, (r) #548, (s) #568, (t) #618, (u) #678, (v) #718, (w) #748, and (x) #766.

per second. While in the MFMC algorithm, we iterate the mean field fix-point equation six times and the mean values of the recovered mean field distribution are adopted as the the tracking result. It can roughly run at the speed of 0.6 frames per second, just similar to what has been reported in [7]. We also use 200 particles for each of the body parts.

Another issue of interest would be that if using one control parameter T for all the different component of the state random variable $\mathbf{x}_i$ is a good setting. In theory, it will have no problem, but in real experiments, it may encounter problems since different components of $\mathbf{x}_i$ may have different ranges. For example, in the six-dimensional affine motion space, the translation component and the scaling

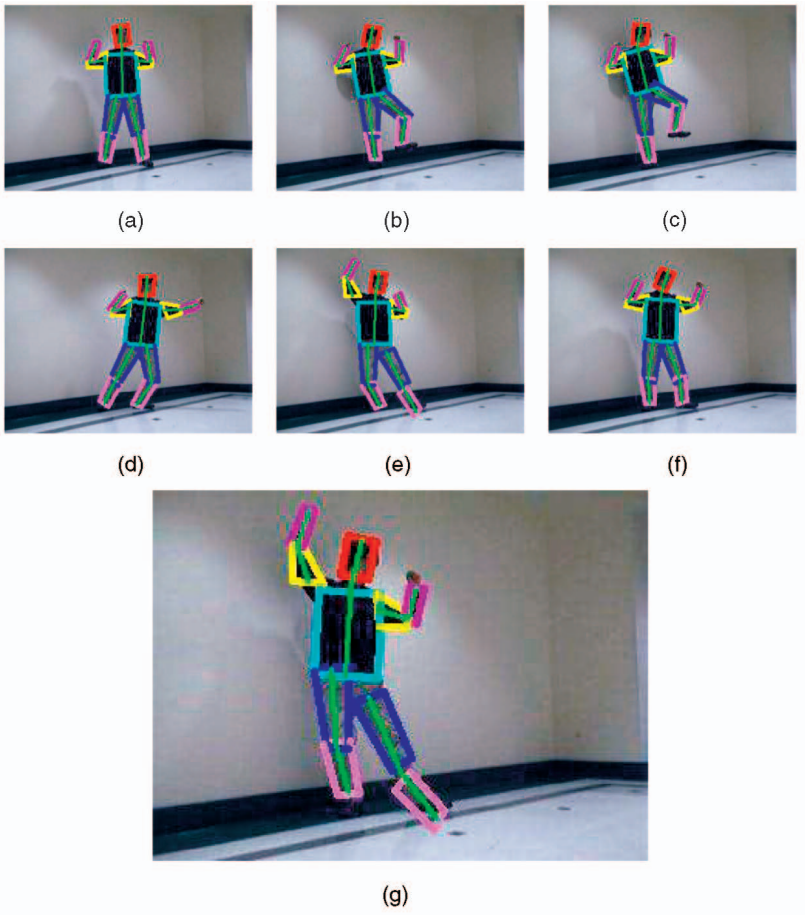

Fig. 10. Tracking articulated body motion by MFMC, the algorithm failed after frame $\#368$ due to the heavy multimodality in the motion posteriors where the mean estimate deviated a lot from the true motion. (a) $\#8$, (b) $\#128$, (c) $\#152$, (d) $\#208$, (e) $\#143$, (f) $\#308$, and (g) $\#370$.

component have different ranges. Thus, we design different annealing schemes for different component of the affine state vector, i.e., the annealing of the translation components of $\mathbf{x}_i$ starts at $\mathbf{T}_{max1} = 8$ and the annealing of the scaling components of $\mathbf{x}_i$ starts at $\mathbf{T}_{max2} = 0.6$.

8 Conclusion and Future Work

This paper proposed a novel variational MAP algorithm for the optimal MAP estimation of complex stochastic systems. By constraining the mean field variational distribution to be multivariate Gaussian, a DA scheme is naturally incorporated into the mean field variational analysis to pursue the optimal MAP estimation. Our main contributions are:

1. We show that the limit of the topology of the KL divergence between a multivariate Gaussian distribution $g(\mathbf{X}) = \mathcal{N}(\mathbf{X}|\vec{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2\mathcal{I})$ and an arbitrary p.d.f. $p(\mathbf{X})$, when the $\boldsymbol{\sigma}^2$ approaches to zero, will converge to the topology of $-log(\vec{\boldsymbol{\mu}})$ (see Lemma 2 in Appendix 1). Thus, there is an one-to-one correspondence of the minima between the $\lim_{\boldsymbol{\sigma}^2 \to \mathbf{0}} KL(g(\mathbf{X})||p(\mathbf{X}))$ and the maxima of the p.d.f. $p(\mathbf{X})$, and the limit of the infimum point of the KL divergence will converge to the supreme point of the $p(\mathbf{X})$, as shown in Theorem 1.
2. Based on Theorem 1, we nicely incorporate a DA scheme into the Gaussian constrained mean field variational analysis to pursue the optimal MAP estimation of complex stochastic systems. Although DA may not guarantee global optimality, our extensive synthetic and real experiments show that it is very likely to achieve a global or near global optimal result. Therefore, we achieve an efficient and effective way for optimal MAP estimation.

There are also several questions need to be further investigated:

1. Although we have empirically shown that the mean field fix-point iteration in (19) and (20) is superior to the iteration in (23) and (24), i.e., the latter two failed to converge even in a relative simple synthetic problem, we are still interested in investigating theoretically why the former two equations can obtain better results under the context of optimization.
2. Is there an optimal annealing scheme which can guarantee to achieve the optimal results more efficiently? The answer of this question will also reveal the convergence rate of the annealing scheme.
3. When will the proposed variational MAP algorithm achieve the global optimality? Although generally in our experiments, the variational MAP algorithm with the hyperbolic decreasing DA scheme achieves good results, practically there is no guarantee that

the algorithm will achieve global optimality. Should there be a sufficient condition, or a necessary condition, or both for the global optimality?

4. With regards to applying the proposed variational MAP algorithm to computer vision problems, should there be an efficient way of incorporating some bottom-up processing to facilitate more efficient convergence of the algorithm? The answer of this question will achieve a data driven variational MAP algorithm for many computer vision problems.

We will further study the above questions in our future work.

APPENDIX 1

LEMMA OF THEOREM 1

Define, for $\boldsymbol{\sigma} > 0$, the quantity

$$\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}) = E_q\{\log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x})\} = \int_{\mathbf{x}} q(\mathbf{x}) \log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x}. \quad (32)$$

Note that (2) ensures that $\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}})$ is finite provided that σ is small enough, as we will show in the proof of Lemma 2. We propose the following lemmas to facilitate the proof of Theorem 1.

Lemma 1. *Under the same conditions of Theorem 1, we have*

$$KL(q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) \| p(\mathbf{x})) = \mathcal{C}_{\boldsymbol{\sigma}} - \varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}), \quad (33)$$

where $\mathcal{C}_{\boldsymbol{\sigma}}$ is a constant relied only on $\boldsymbol{\sigma}$.

Proof.

$$\begin{aligned}
KL(q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) \| p(\mathbf{x})) &= \int_{\mathbf{x}} q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) \log \frac{q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} && (34)\\
&= \int_{\mathbf{x}} q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) \log q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) d\mathbf{x} - \int_{\mathbf{x}} q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} && (35)\\
&= -\log\{(2\pi e)^n \boldsymbol{\sigma}^n\} - \int_{\mathbf{x}} q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} && (36)\\
&= \mathcal{C}_{\boldsymbol{\sigma}} - \int_{\mathbf{x}} q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} && (37)\\
&= \mathcal{C}_{\boldsymbol{\sigma}} - \int_{\mathbf{x}} \boldsymbol{\sigma}^n q^{\vec{\boldsymbol{\mu}}}_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) \log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} && (38)\\
&= \mathcal{C}_{\boldsymbol{\sigma}} - \int_{\mathbf{x}} q(\mathbf{x}) \log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} && (39)\\
&= \mathcal{C}_{\boldsymbol{\sigma}} - E_q\{log(p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}))\} && (40)\\
&= \mathcal{C}_{\boldsymbol{\sigma}} - \varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}). && (41)
\end{aligned}$$

□

Lemma 2. *Under the same conditions of Theorem 1, we have that for any $\boldsymbol{\mu} \in \mathcal{R}^n$,*

$$\lim_{\boldsymbol{\sigma} \to 0} \varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}) = \log p(\vec{\boldsymbol{\mu}}), \quad (42)$$

Proof. First, (2) guarantees that, for $\boldsymbol{\sigma}$ sufficiently small, $\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}})$ is finite for all $\vec{\boldsymbol{\mu}} \in \mathcal{R}^n$, i.e., if $\boldsymbol{\sigma} < \frac{\sqrt{2}}{2}$, note by parallelogram law $-(\mathbf{x} - \vec{\boldsymbol{\mu}})^T(\mathbf{x} - \vec{\boldsymbol{\mu}}) \le \vec{\boldsymbol{\mu}}^T\vec{\boldsymbol{\mu}} - \frac{\mathbf{x}^T\mathbf{x}}{2}$, we have

$$\begin{aligned}
|\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}})| &= \left| \int_{\mathbf{x}} q(\mathbf{x}) \log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} \right| && (43)\\
&= \left| \int_{\mathbf{x}} \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right) \log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} \right| && (44)\\
&\le \int_{\mathbf{x}} \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right) |\log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x})| d\mathbf{x} && (45)\\
&= \int_{\mathbf{x}} \frac{1}{(2\pi)^{\frac{n}{2}}\boldsymbol{\sigma}^n} \exp\left(-\frac{(\mathbf{x} - \vec{\boldsymbol{\mu}})^T(\mathbf{x} - \vec{\boldsymbol{\mu}})}{2\boldsymbol{\sigma}^2}\right) |\log p(\mathbf{x})| d\mathbf{x} && (46)\\
&\le \int_{\mathbf{x}} \frac{1}{(2\pi)^{\frac{n}{2}}\boldsymbol{\sigma}^n} \exp\left(-(\mathbf{x} - \vec{\boldsymbol{\mu}})^T(\mathbf{x} - \vec{\boldsymbol{\mu}})\right) |\log p(\mathbf{x})| d\mathbf{x} && (47)\\
&\le \int_{\mathbf{x}} \frac{1}{(2\pi)^{\frac{n}{2}}\boldsymbol{\sigma}^n} \exp\left(\vec{\boldsymbol{\mu}}^T\vec{\boldsymbol{\mu}} - \frac{\mathbf{x}^T\mathbf{x}}{2}\right) |\log p(\mathbf{x})| d\mathbf{x} && (48)\\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\boldsymbol{\sigma}^n} \exp\left(\vec{\boldsymbol{\mu}}^T\vec{\boldsymbol{\mu}}\right) \int_{\mathbf{x}} \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right) |\log p(\mathbf{x})| d\mathbf{x} && (49)\\
&< +\infty. && (50)
\end{aligned}$$

By the continuity of $p(\mathbf{x})$, for any $\epsilon > 0$, there exists $\delta = \delta(\vec{\boldsymbol{\mu}}) > 0$ such that $|\mathbf{x} - \vec{\boldsymbol{\mu}}| < \delta$ implies $|\log p(\mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})| < \epsilon$, we have

$$\begin{aligned}
&|\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}) - \log p(\vec{\boldsymbol{\mu}})|\\
&= \left| \int_{\mathbf{x}} q(\mathbf{x}) \log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} - \log p(\vec{\boldsymbol{\mu}}) \right| && (51)\\
&= \left| \int_{\mathbf{x}} q(\mathbf{x}) (\log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})) d\mathbf{x} \right| && (52)\\
&= \left| \int_{|\mathbf{x}| < \frac{\delta}{\boldsymbol{\sigma}}} q(\mathbf{x}) (\log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})) d\mathbf{x} \right. && (53)\\
&\left. + \int_{|\mathbf{x}| > \frac{\delta}{\boldsymbol{\sigma}}} q(\mathbf{x}) (\log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})) d\mathbf{x} \right| && (53)\\
&\le \left| \epsilon \int_{|\mathbf{x}| < \frac{\delta}{\boldsymbol{\sigma}}} q(\mathbf{x}) d\mathbf{x} \right| + \left| \int_{|\mathbf{x}| > \frac{\delta}{\boldsymbol{\sigma}}} q(\mathbf{x}) (\log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})) d\mathbf{x} \right| && (54)\\
&= \epsilon P\left(|\mathbf{x}_q| < \frac{\delta}{\boldsymbol{\sigma}}\right) + \left| \int_{|\mathbf{x}| > \frac{\delta}{\boldsymbol{\sigma}}} q(\mathbf{x}) (\log p(\vec{\boldsymbol{\mu}} + \boldsymbol{\sigma}\mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})) d\mathbf{x} \right| && (55)\\
&< \epsilon + \int_{|\mathbf{x}| > \delta} \boldsymbol{\sigma}^{-n} q\left(\frac{\mathbf{x}}{\boldsymbol{\sigma}}\right) |\log p(\vec{\boldsymbol{\mu}} + \mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})| d\mathbf{x}. && (56)
\end{aligned}$$

For the second term in (56), we have shown from (43) to (50) that it is integrable. Moreover, for $\boldsymbol{\sigma}$ small enough and for any fixed $|\mathbf{x}| > \delta$, it is easy to show that

$$\lim_{\boldsymbol{\sigma} \to 0} \boldsymbol{\sigma}^{-n} q\left(\frac{\mathbf{x}}{\sigma}\right) = 0. \quad (57)$$

Therefore, for any fixed $|\mathbf{x}| > \delta$ and for $\boldsymbol{\sigma}$ small enough, we have

$$\boldsymbol{\sigma}^{-n} q\left(\frac{\mathbf{x}}{\sigma}\right) < \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right), \quad (58)$$

thus, it follows that

$$\begin{aligned}
&\boldsymbol{\sigma}^{-n} q\left(\frac{\mathbf{x}}{\sigma}\right) |\log p(\vec{\boldsymbol{\mu}} + \mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})|\\
&\quad < \exp\left(-\frac{\mathbf{x}^T\mathbf{x}}{2}\right) |\log p(\vec{\boldsymbol{\mu}} + \mathbf{x}) - \log p(\vec{\boldsymbol{\mu}})|.
\end{aligned} \quad (59)$$

From the integrability condition in (2) and the properness of $p(\mathbf{x})$, it is easy to figure out that the right hand side of (59) is also integrable. With (57), applying Lebesgue's dominated convergence theorem, we have that the second term in (56) goes to zero as $\boldsymbol{\sigma} \to 0$, i.e., for the given $\epsilon > 0$, there exists a $\boldsymbol{\sigma}_1 = \boldsymbol{\sigma}(\epsilon) > 0$ such that when $\boldsymbol{\sigma} < \boldsymbol{\sigma}_1$,

$$|\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}) - \log p(\vec{\boldsymbol{\mu}})| < \epsilon + \epsilon = 2\epsilon. \tag{60}$$

Then, we have

$$\lim_{\boldsymbol{\sigma}\to 0} |\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}) - \log p(\vec{\boldsymbol{\mu}})| = 0. \tag{61}$$

Therefore, (42) holds. □

Lemma 3. *Under the same condition of Theorem 1, if a sequence* $\{\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}\}$ *is such that*

$$\lim_{\boldsymbol{\sigma}\to 0} \varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}) = \sup_{\mathbf{x}} \log p(\mathbf{x}), \tag{62}$$

then

$$\lim_{\boldsymbol{\sigma}\to 0} \vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} = \mathbf{x}^*. \tag{63}$$

Proof. We need to prove that, if $\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}) \to \log p(\mathbf{x}^*)$ as $\boldsymbol{\sigma} \to 0$, it is impossible that there exists $\delta > 0$ such that infinitely often $|\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} - \mathbf{x}^*| \geq 2\delta$. Let us first assume that such a δ exists, then according to the continuity of $\log p(\mathbf{x})$, there exists a $\epsilon > 0$ such that $\log p(\mathbf{x}) < \log p(\mathbf{x}^*) - \epsilon$ for $|\mathbf{x} - \mathbf{x}^*| \geq \delta$. For $\boldsymbol{\sigma}$ small enough, e.g., $\boldsymbol{\sigma} < \frac{1}{3}\delta$, we then have

$$\begin{aligned}
\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}) &= \int_{\mathbf{x}} q(\mathbf{x}) \log p(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} && (64)\\
&= \int_{|\mathbf{x}|<\frac{\delta}{\boldsymbol{\sigma}}} q(\mathbf{x}) \log p(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} \\
&\quad + \int_{|\mathbf{x}|>\frac{\delta}{\boldsymbol{\sigma}}} q(\mathbf{x}) \log p(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} && (65)\\
&\leq (\log p(\mathbf{x}^*) - \epsilon) P\left(\frac{|\mathbf{x}_q| < \delta}{\boldsymbol{\sigma}}\right) \\
&\quad + \log p(\mathbf{x}^*) P\left(|\mathbf{x}_q| \geq \frac{\delta}{\boldsymbol{\sigma}}\right) && (66)\\
&= \log p(\mathbf{x}^*) - \epsilon P\left(|\mathbf{x}_q| < \frac{\delta}{\boldsymbol{\sigma}}\right) && (67)\\
&\leq \log p(\mathbf{x}^*) - \frac{1}{2}\epsilon && (68)\\
&\Longrightarrow \\
&|\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}) - \log p(\mathbf{x}^*)| \geq \frac{1}{2}\epsilon, && (69)
\end{aligned}$$

where (66) holds because for $|\mathbf{x}| < \frac{\delta}{\boldsymbol{\sigma}}$, we have $|\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} + \boldsymbol{\sigma}\mathbf{x} - \mathbf{x}^*| \geq \delta$ which implies that $\log p(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} + \boldsymbol{\sigma}\mathbf{x}) < \log p(\mathbf{x}^*) - \epsilon$.

Equation (69) immediately contradicts with (62). Therefore, (63) holds given that (62) holds. □

Appendix 2
Proof of Theorem 1

Proof. We first proceed to prove

$$\lim_{\boldsymbol{\sigma}\to 0} \sup_{\vec{\boldsymbol{\mu}}} \varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}) = \sup_{\mathbf{x}} \log p(\mathbf{x}). \tag{70}$$

Note that from Lemma 1, the series $\{\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}\}$ in Theorem 1 is also such that

$$\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}) = \sup_{\vec{\boldsymbol{\mu}}} \varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}). \tag{71}$$

First of all, we have $\log p(\mathbf{x}) < \log p(\mathbf{x}^*) + \epsilon$ for any $\epsilon > 0$. Thus, for any $\boldsymbol{\sigma} > 0$, we have

$$\begin{aligned}
\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}) &= \int_{\mathbf{x}} q(\mathbf{x}) \log p(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}} + \boldsymbol{\sigma}\mathbf{x}) d\mathbf{x} && (72)\\
&< (\log p(\mathbf{x}^*) + \epsilon) \int_{\mathbf{x}} q(\mathbf{x}) d\mathbf{x} && (73)\\
&= \log p(\mathbf{x}^*) + \epsilon. && (74)
\end{aligned}$$

Moreover, from Lemma 2, we have, for any $\epsilon > 0$, there exists a $\boldsymbol{\sigma}_1 > 0$, for $\boldsymbol{\sigma} < \boldsymbol{\sigma}_1$, we have

$$|\varphi_{\boldsymbol{\sigma}}(\mathbf{x}^*) - \log p(\mathbf{x}^*)| < \epsilon. \tag{75}$$

Then, from (74)} and (75), we easily obtain that for $\boldsymbol{\sigma} < \boldsymbol{\sigma}_1$,

$$\log p(\mathbf{x}^*) - \epsilon < \varphi_{\boldsymbol{\sigma}}(\mathbf{x}^*) < \varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}) < \log p(\mathbf{x}^*) + \epsilon. \tag{76}$$

Thus, for any $\epsilon > 0$, there exists a $\boldsymbol{\sigma}_1 > 0$, for $\boldsymbol{\sigma} < \boldsymbol{\sigma}_1$, we have

$$|\varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}_{\boldsymbol{\sigma}}) - \log p(\mathbf{x}^*)| = \left| \sup_{\vec{\boldsymbol{\mu}}} \varphi_{\boldsymbol{\sigma}}(\vec{\boldsymbol{\mu}}) - \sup_{\mathbf{x}} \log p(\mathbf{x}) \right| < \epsilon. \tag{77}$$

This immediately proves (70). Then, we can directly conclude that (3) holds by applying Lemma 3. □

Acknowledgments

This work was supported in part by US National Science Foundation Grants IIS-0347877, IIS-0308222, and Northwestern faculty startup funds for Ying Wu and Walter P. Murphy Fellowship for Gang Hua.

References

[1] C. Andrieu and A. Doucet, "Joint Bayesian Model Selection and Esitimation of Noisy Sinusoids via Reversible Jump MCMC," *IEEE Trans. Signal Processing*, vol. 47, no. 10, pp. 2667-2676, 1999.
[2] Z. Tu and S.-C. Zhu, "Image Segmentation by Data-Driven Markov Chain Monte Carlo," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 657-673, May 2002.
[3] A. Barbu and S.-C. Zhu, "Graph Partition by Swendsen-Wang Cut," *Proc. IEEE Int'l Conf. Computer Vision*, 2003.
[4] W.T. Freeman and E.C. Pasztor, "Learning Low-Level Vision," *Proc. IEEE Int'l Conf. Computer Vision*, pp. 1182-1189, 1999.
[5] W.T. Freeman and E.C. Pasztor, "Markov Network for Low-Level Vision," technical report, MERL, Mitsubish Electric Research Laboratory, 1999.
[6] Y. Wu and T.S. Huang, "Robust Visual Tracking by Co-Inference Learning," *Proc. IEEE Int'l Conf. Computer Vision*, vol. 2, pp. 26-33, 2001.

[7] Y. Wu, G. Hua, and T. Yu, "Tracking Articulated Body by Dynamic Markov Network," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1094-1101, 2003.

[8] L. Sigal, M. Isard, B. Sigelman, and M. Black, "Attractive People: Assembling Loose-Limbed Models Using Non-Parametric Belief Propagation," *Advances in Neural Information Processing System 16,* MIT Press, 2004.

[9] J. Yedidia, W. Freeman, and Y. Weiss, "Understanding Belief Propagation and Its Generalization," *Exploring Artificial Intelligence in the New Millenium,* chapter 8, pp. 239-286, Elsevier Science and Technology Books, 2003.

[10] G. Hua, Y. Wu, and T. Yu, "Analyzing Structured Deformable Shapes via Mean Field Monte Carlo," *Proc. IEEE Asia Conf. Computer Vision,* 2004.

[11] M.J. Beal, "Variational Algorithms for Approximate Bayesian Inference," PhD Thesis, Gatsby Computational Neuroscience Unit, Univ. College, London, 2003.

[12] J.M. Winn, "Variational Message Passing and Its Application," PhD thesis, Dept. of Physics, Univ. of Cambridge, 2003.

[13] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, "Equations of State Calculations by Fast Computing Machine," *J. Chemical Physics,* vol. 21, pp. 1087-1091, 1953.

[14] J.M.P.V.S. Kirkpatrick and C.D. Gelatt, "Optimization by Simulated Annealing," *Science,* vol. 220, no. 4598, pp. 671-680, 1983.

[15] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 6, pp. 721-741, June 1984.

[16] A.L. Yuille and J.J. Kosowsky, "Statistical Physics Algorithms that Converge," *Neural Computation,* vol. 6, no. 3, pp. 341-356, June 1994.

[17] J. Puzicha, T. Hofmann, and J.M. Buhmann, "Deterministic Annealing: Fast Physical Heuristics for Real-Time Optimization of Large Systems," *Proc. 15th IMACS World Conf. Scientific Computation, Modelling and Applied Math.,* 1997.

[18] J. Puzicha and J.M. Buhmann, "Multiscale Annealing for Grouping and Unsupervised Texture Segmentation," *Computer Vision and Image Understanding (CVIU),* vol. 76, no. 3, pp. 213-230, 1999.

[19] S.Z. Li, "Robustizing Robust M-Estimation Using Deterministic Annealing," *Pattern Recognition,* vol. 29, no. 1, pp. 159-166, 1996.

[20] K.P. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning," PhD thesis, Computer Science Division, Univ. of California, Berkeley, 2002.

[21] V.I. Pavlovic, "Dynamic Bayesian Networks for Information Fusion with Application to Human-Computer Interfaces," PhD thesis, Dept. of Electrical and Computer Eng., Univ. of Illinois at Urbana-Champaign, 1999.

[22] G. Hua and Y. Wu, "Multi-Scale Visual Tracking by Sequential Belief Propagation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 826-833, 2004.

[23] M. Jordan and Y. Weiss, "Graphical Models: Probabilistic Inference," *The Handbook of Brain Theory and Neural Network,* second ed. MIT Press, pp. 243-266, 2002.

[24] Y. Wang, T. Tan, and K.-F. Loe, "Video Segmentation Based on Graphical Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 335-342, 2003.

[25] Y. Wu, G. Hua, and T. Yu, "Switching Observation Models for Contour Tracking in Clutter," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 295-302, 2003.

[26] Y. Wu, T. Yu, and G. Hua, "Tracking Appearances with Occlusions," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 789-795, 2003.

[27] K. Murphy, Y. Weiss, and M. Jordan, "Loopy-Belief Propagation for Approximate Inference: An Empirical Study," *Proc. 15th Conf. Uncertainty in Artificial Intelligence,* 1999.

[28] W.T. Freeman and H. Zhang, "Shape-Time Photography," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003.

[29] A. Blake and M. Isard, *Active Contours.* Springer-Verlag, 1998.

[30] M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of Conditional Density," *Proc. European Conf. Computer Vision,* vol. 1, pp. 343-356, 1996.

[31] M. Isard and A. Blake, "Condensation-Conditional Density Propagation for Visual Tracking," *Int'l J. Computer Vision,* vol. 29, no. 1, pp. 5-28, 1998.

[32] T.S. Jaakkola, "Tutorial on Variational Approximation Method," *Advanced Mean Field Methods: Theory and Practice,* MIT Press, 2000.

[33] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky, "Nonparametric Belief Propagation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 605-612, 2003.

[34] M. Isard, "Pampas: Real-Valued Graphical Models for Computer Vision," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 613-620, 2003.

[35] A.V. Rao, D.J. Miller, K. Rose, and A. Gersho, "A Deterministic Annealing Approach for Parsimonious Design of Piecewise Regression Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 2, pp. 159-173, Feb. 1999.

[36] D. Doll and W. von Seelen, "Object Recognition by Deterministic Annealing," *Image and Vision Computing,* vol. 15, pp. 855-860, 1997.

[37] J. Deutscher, A. Blake, and I. Reid, "Articulated Body Motion Capture by Annealed Particle Filtering," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2000.

[38] T.M. Cover and J.A. Thomas, *Elements of Information Theory,* D.L. Schilling, ed., second ed. New York: John Wiley and Sons, 1991.

[39] J.B. Rosen, "The Gradient Projection Method for Nonlinear Programming. Part I. Linear Constraints," *J. Soc. Industrial and Applied Math.,* vol. 8, no. 1, pp. 181-217, Mar. 1960.

[40] J.B. Rosen, "The Gradient Projection Method for Nonlinear Programming. Part II. Nonlinear Constraints," *J. Soc. Industrial and Applied Math.,* vol. 9, no. 4, pp. 514-532, Dec. 1961.

Gang Hua received the MS degree in control science and engineering at XJTU in 2002. he is a PhD candidate in the Department of Electrical and Computer Engineering, Northwestern University. He has been working with Professor Ying Wu since September 2002. His main research interests include computer vision, computer graphics, and machine learning. Before attending Northwestern, he was a master student in the AI&R Institute at Xi'an Jiaotong University (XJTU), Xi'an, People's Republic of China, under the supervision of Professor Nanning Zheng. He was enrolled in the Special Class for the Gifted Young of XJTU in 1994 and received the BS degree in automatic control engineering at XJTU in 1999. He received the Walter P. Murphy Fellowship at Northwestern University in 2002. When he was at XJTU, he was awarded the Guanghua Fellowship, the EastCom Research Scholarship, the Most Outstanding Student Exemplar Fellowship, the Sea-star Fellowship, and the Jiangyue Fellowship in 2001, 2000, 1997, 1997, and 1995, respectively. He was also a recipient of the University Fellowship for Outstanding Student at XJTU from 1994 to 2002. He is a student member of the IEEE.

Ying Wu received the BS degree from the Huazhong University of Science and Technology, Wuhan, China, in 1994, the MS degree from Tsinghua University, Beijing, China, in 1997, and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC), Urbana, Illinois, in 2001. From 1997 to 2001, he was a graduate research assistant at the Image Formation and Processing Group of the Beckman Institute for Advanced Science and Technology at UIUC. During the summers of 1999 and 2000, he was a research intern with the Vision Technology Group, Microsoft Research, Redmond, Washington. Since 2001, he has been on the faculty of the Department of Electrical and Computer Engineering at the Northwestern University, Evanston, Illinois. His current research interests include computer vision, computer graphics, machine learning, human-computer intelligent interaction, image/video processing, multimedia, and virtual environments. He received the Robert T. Chien Award at the University of Illinois at Urbana-Champaign in 2001, and is a recipient of the US National Science Foundation CAREER award. He is a member of the IEEE and the IEEE Computer Society.