Abstract of "Variational Approximations with Diverse Applications" by Jason L. Pacheco, Brown University Computer Science, March 2016.

We develop a family of algorithms for statistical inference in models of high dimensional continuous random variables. Our approach builds on existing variational methods, which provide computationally efficient alternatives to Markov chain Monte Carlo (MCMC) sampling. While efficient, existing variational approximations are not applicable to many continuous models of practical interest, or they are altogether unstable and produce degenerate solutions. We construct a more powerful class of algorithms for posterior marginal and maximum a posteriori (MAP) inference which avoids these limitations. Throughout this thesis we present a series of vignettes demonstrating the generality of our approach on disparate applications such as: human pose estimation in images and video, protein structure prediction, and target tracking.

We begin by considering MAP inference problems for continuous Markov random fields (MRFs) where the well-known max-product (MP) variant of belief propagation (BP) cannot be applied due to non-Gaussian statistics. Motivated by similar ideas in sum-product BP we develop a particle-based approximation of the continuous MP messages. Unique to the MAP setting, however, is a need for diversity among the hypotheses, to avoid classic particle degeneracies. Using an integer programming formulation we enforce particle diversity, from which we can recover a set of distinct local maxima. Our nonparametric approximation applies to any model for which the probability density can be evaluated in a black-box manner, even for models with no analytic form. We validate our approach using a model for estimating human pose from single images and videos.

To further motivate and validate our approach we consider the challenging problem of estimating three-dimensional protein structures. Using our particle-based approximations we optimize the continuous energy function encoding protein stability, and thereby avoid discrete approximations employed by most existing methods. In this way we are able to recover fine details of protein structure which standard methods fail to capture, and by preserving diverse hypotheses our approach maintains the conformational diversity proteins are known to exhibit.

The final component of this thesis explores variational methods for posterior marginal inference. We begin by developing inference based on expectation propagation (EP) for tracking a time-evolving target in the presence of measurement error and clutter detections. Our method outperforms existing tracking algorithms

while generalizing classical techniques. Motivated by non-convergence and degeneracy issues that are observed in this setting, we formulate a convergent nonlinear optimization which uses an augmented Lagrangian technique with provable convergence guarantees. Moreover, we identify the set of constraints that, when violated, produce unnormalizable marginal approximations in message passing fixed points. Using gradient projection we strictly enforce these normalization constraints to guarantee variational approximations are well-formed. These techniques outperform loopy BP and EP on MRFs with discrete, Gaussian and Gaussian mixture distributions.

Variational Approximations with Diverse Applications

by
Jason L. Pacheco
B. S., U. Massachusetts Dartmouth, 2003
Sc. M., Brown University, 2008

A dissertation submitted in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy
in the Department of Computer Science at Brown University

Providence, Rhode Island
March 2016

This dissertation by Jason L. Pacheco is accepted in its present form by the Department of Computer Science as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Erik Sudderth, Advisor

Recommended to the Graduate Council

Date _____ _____

Michael Littman, Reader

Date _____ _____

Alexander Ihler, Reader
Dept. of Computer Science, UC Irvine

Approved by the Graduate Council

Date _____ _____

Dean of the Graduate School

Contents

Chapter 1

Introduction

Graphical models are used to express complex global relationships by specifying simpler local interactions. In computer vision, for example, graphical models of human pose are defined via a *loose-limbed* approach where neighboring parts are joined by *springs* and orientation is given by relative displacement and rotation [151, 50]. While the model is defined by simple pairwise relationships, it is sufficiently expressive to capture complex global variations in pose and appearance. In other disciplines, such as computational biology, the complex 3D structure of a protein molecule can be modeled by simple pairwise energetic relationships between groups of atoms [137]. However, a model alone is not useful without the ability to perform inferences or estimate unknown quantities based on observed information.

It is precisely the expressiveness of graphical models that makes statistical inference so difficult. For example, the Ising lattice from statistical physics encodes the, sometimes chaotic, dynamics of interacting electrons, such as phase transitions, or the dynamics of disordered magnets known as *spin glasses*. Though the model definition is simple, the chaotic global dynamics it encodes make estimating the unknown electron states computationally intractable [149].

The most challenging inference problems arise in continuous models with high-dimensional non-Gaussian statistics. In this thesis we develop algorithms for approximate marginal and maximum a posteriori inference for such models, where existing techniques do not apply. We focus on algorithms that apply to a broad class of models, regardless of their analytic form, and explore these algorithms in a variety of contexts.

1.1 Finding Modes of Continuous Distributions

In many applications the natural inference task is to estimate the most likely configuration of unknowns, given observed data. So-called *maximum a posteriori* (MAP) inference is particularly important for models of physical systems, such as the human body or protein molecules. Physical constraints imposed by these models prohibit arbitrary global configurations, and so it is important to reason about jointly consistent solutions that are feasible.

For continuous models MAP inference reduces to a nonlinear optimization, often with many local optima due to the complexity of the underlying distribution. Moreover, global optima are not always preferable due to inaccuracies in model specification. To be robust to model mismatch it is important to capture multiple local optima, since they often correspond to good solutions. In models of protein structure for example, where form is closely linked to function, it is known that protein molecules assume multiple stable configurations, and characterizing these is important [165, 100, 109].

In this thesis we develop *particle-based* MAP inference for the so-called *diverse M-best MAP* inference problem [17]. Our approach is inspired by a similar method for approximate marginal inference [79], but incorporates a notion of *diversity* in the set of hypotheses to capture multiple local optima. This nonparametric approximation enables MAP inference for an arbitrary probability distribution without imposing analytic restrictions on the model class. While the framework is entirely general, we focus on applications involving articulated physical models of human bodies and protein structure.

1.1.1 Human Pose Estimation and Tracking

Analysis of images involving human figures (sometimes called "looking at people") has continued to be one of the most active application areas in computer vision for more than a decade. Reasoning about human behavior from images and video involves problems such as detection, localization and tracking [122, 75, 51]. Accurately reasoning about human pose provides an informative cue for automated systems, e.g. for image understanding, activity recognition, and "smart" surveillance [156, 185, 61].

Models of human pose largely began with the Pictorial Structures (PS) of Fischler and Elschlager [53] and later refined by Felzenszwalb and Huttenlocher [50, 51]. The PS model represents articulated limbs by rectangular bounding boxes, capturing only high-level shape information such as length and area. The "cardboard people" models of Ju et al. [89] attempt to capture minor shape variations with polygonal regions, while a richer shape model is expressed in terms of edge contours by Freifeld et al. [55].

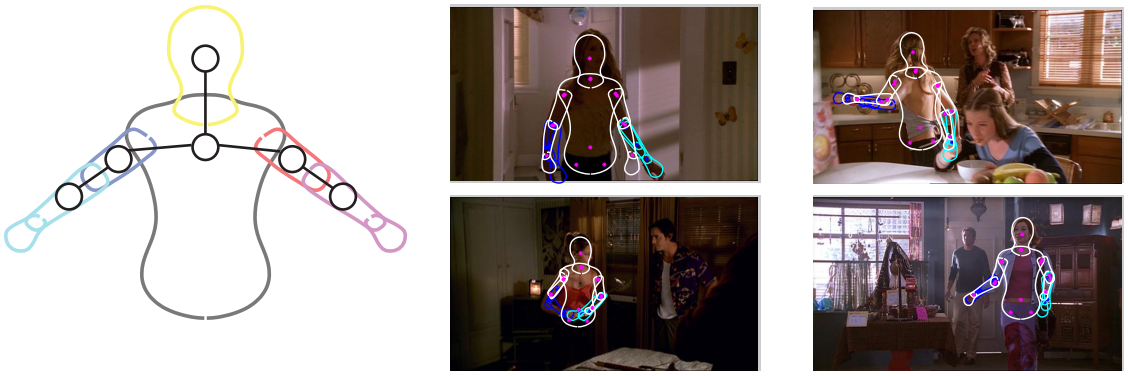


Figure 1.1: **Human pose estimation.** *Left:* Deformable structures model of human pose and shape. *Right:* Example estimates of pose and shape on the Buffy dataset.

The Deformable Structures (DS) model of Zuffi et al. [187] combines this notion of smooth shape deformations with the PS model by replacing bounding boxes with a PCA shape model.

Despite advancements in modeling the statistics of human pose, reliable inference remains a challenge. Articulated pose models represent the orientation of parts by their relative displacement and rotations. These representations result in high-dimensional energy minimization problems for which existing methods can often prove unreliable. The PS model, for example, expresses energy in terms of local potentials for each part which encode image evidence, and pairwise spring-like potentials that constrain deformations between neighboring parts. This minimization is non-trivial and Felzenszwalb and Huttenlocher [50] suggest a dynamic programming approach based on a discretized state-space. Variations on the basic PS model with discrete max-product are numerous [8, 24, 48, 133, 134], and in particular Yang and Ramanan [176] have produced state-of-the-art results on one evaluation by learning a mixture representation of part likelihoods.

Building on the DS model of Zuffi et al. [187] we express human pose estimation as a continuous energy minimization. By also modeling shape variation the problem size is drastically increased, further necessitating inference which can reliably operate in a high-dimensional continuous space.

Estimating human pose from video sequences extends the pose estimation task temporally. Ferrari et al. [52] apply the PS model to pose tracking while developing progressive search-space reduction techniques to allow for discretization of the high-dimensional solution space. The authors find no benefit in performing inference on the full temporal model, indeed they show that results degrade when the temporal correlations are incorporated. This finding is echoed by Sapp et al. [139] where the authors also incorporate optical flow as a temporal cue. Zuffi et al. [188] exploit

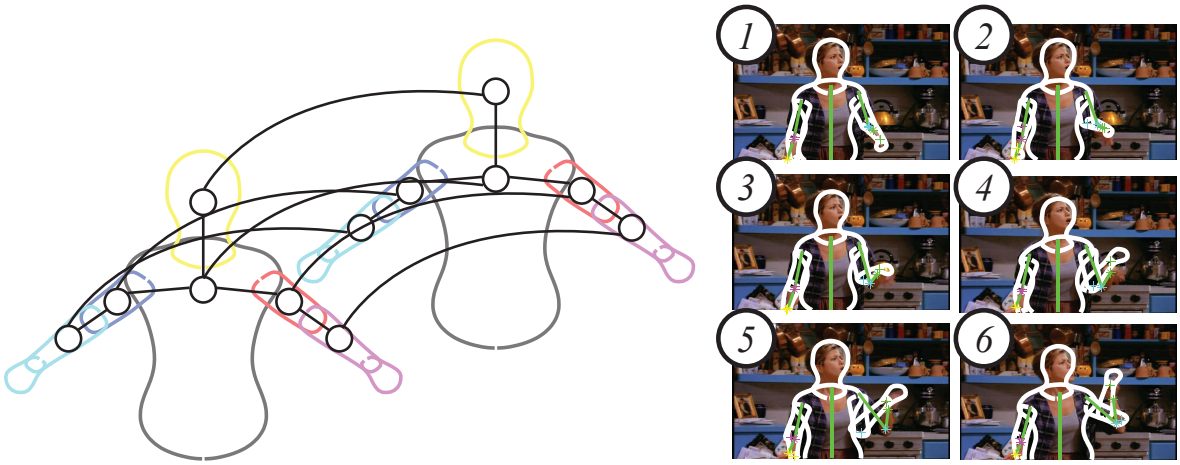


Figure 1.2: **Human pose tracking.** *Left:* Deformable structures model with temporal extensions. *Right:* Example estimates of pose over a video sequence.

optical flow with their *flowing puppets* model via flow-based proposal distributions within particle swarm optimization. These temporal links, however, are not explicitly modeled and inference is performed on individual frames.

Among the previous works for pose estimation in video sequences it remains to be demonstrated that there is a significant advantage to performing inference on the structural and temporal model, jointly. We strongly suspect this is a result of ineffective inference on the high-dimensional energy function. Pruning and discretization of the solution space is too coarse to be of practical use, while local search, particle filters and particle swarm optimization do not exploit the structure of the model. In Chapter 3 we develop particle-based max-product inference for human pose tracking, which allows for optimization in the continuous space while exploiting the model structure. Extending this to the temporal domain we will incorporate flow-based proposals while performing joint inference on the full joint distribution over structural, as well as temporal components.

1.1.2 Protein Structure Prediction

Proteins comprise a class of macromolecules occurring in the cell and which are necessary for virtually all biological functions. They can act as enzymes, catalyzing biochemical reactions important for cellular tasks such as metabolism. Together with RNA polymerase proteins play a key role in the transcription process where they serve as initiation factors determining where transcription begins on the DNA strand. Proteins even play a major role in creating new proteins, where together with rRNA they comprise the ribosome, the location for protein synthesis.

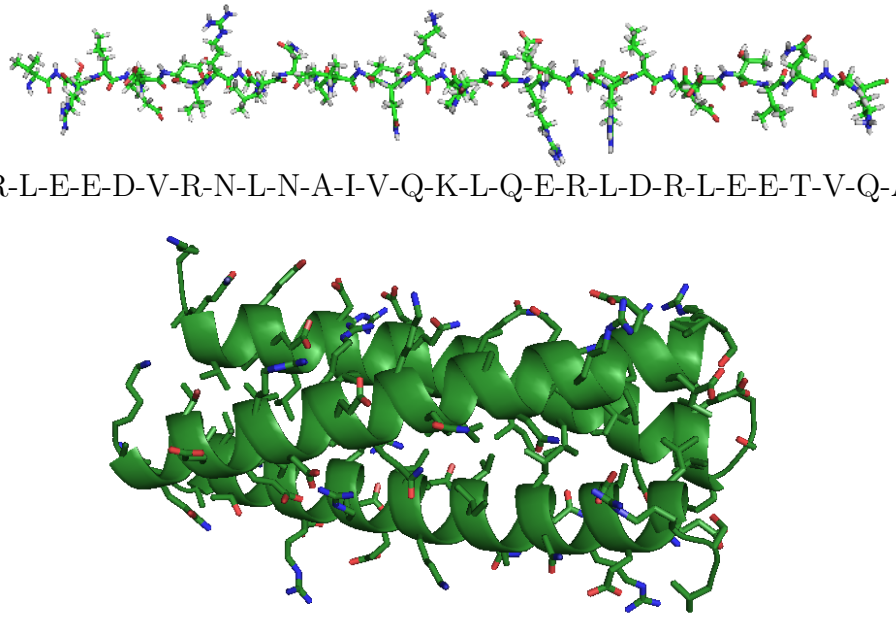


Figure 1.3: **Protein structure.** *Top:* The primary structure given by a sequence of amino acids joined by peptide bonds which form a backbone, and amino acid side chains as sticks. *Bottom:* 3D tertiary structure of protein in its native state. Coils denote secondary structural elements of the backbone, called α-helices. Amino acid side chains are shown as small sticks attached to the backbone.

A protein molecule consists of a chain of amino acids joined by peptide bonds Fig. 1.3 (top). Atomic interactions between nearby amino acids, as well as environmental interactions, cause the protein to assume a three-dimensional structure or a *conformation* Fig. 1.3 (bottom). Predicting this structure is one of the most important tasks in computational microbiology as it determines the binding sites and biological function of the protein.

Protein structure prediction is critical for drug discovery and disease research. The majority of neurodegenerative diseases such as Alzheimer's and amyotrophic lateral sclerosis (ALS) are believed to be linked to misfolded proteins which cause a buildup of insoluble extracellular deposits [143]. While methods exist for experimental validation of a protein's native structure, these methods often involve costly and difficult procedures such as X-ray crystallography and nuclear magnetic resonance (NMR) imaging. Developing more effective prediction algorithms is an active area of research as these would aid experimental validation.

Existing computational methods for structure prediction are primarily limited to Markov chain Monte Carlo sampling combined with simulated annealing to deal with

the highly multimodal energy landscape. Because the problem is extremely high-dimensional, many approaches limit the search space through the use of *fragment assembly* methods which exchange sequences of amino acids for *homologues* of known structure [137]. While fragment assembly performs well experimentally [118], there are many cases where homologues are insufficient or non-existent. In these cases *ab initio* structure prediction estimates the structure of a novel protein by minimizing an energy function based on the physics of atomic interactions [9].

In Chapter 4 we address protein structure prediction using the same particle-based inference as underlying our results in human pose estimation. We show that by minimizing the continuous energy function that models protein structure we are able to avoid inaccuracies associated with the standard discretization. Moreover, our focus on solution diversity preserves multiple stable conformations wknown to be associated with protein function [165, 100, 109].

1.2 Convergent Alternative to Message Passing

Inference algorithms developed in this thesis fall under the class of *variational methods* [169, 85, 88], which pose statistical inference as an optimization problem. The objective function, known as the *variational free energy*, is minimized over the class of probability density functions [181, 183]. Necessary conditions for stationarity are given by the *calculus of variations*, from which variational methods derive their name [62].

In practice the variational free energy is optimized using fixed-point methods known as *message passing* algorithms [169], such as the well-known belief propagation (BP) and expectation propagation (EP) algorithms [127, 114]. While existing message passing algorithms define fixed point iterations corresponding to stationary points of the variational free energy, their greedy dynamics do not distinguish between local minima and maxima, and can fail to converge. For continuous estimation problems, this instability is linked to the creation of invalid marginal estimates, such as Gaussians with negative variance. This behavior is unpredictable and problematic in practice, and leads to uninterpretable approximations.

We instead develop a convergent optimization algorithm which directly minimizes the variational objective while avoiding degenerate marginals. Our approach leverages augmented Lagrangian methods with well-understood convergence properties [21], and uses gradient projection [22] to ensure that marginal approximations are valid at all iterations. We derive general algorithms for discrete and Gaussian pairwise Markov random fields, showing improvements over standard loopy belief propagation. We also apply our method to a hybrid model with both discrete and continuous variables,

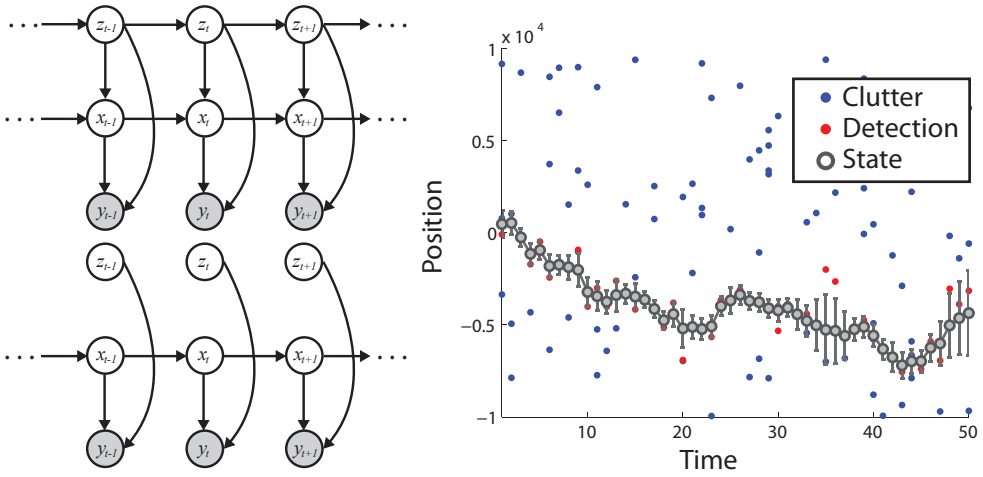


Figure 1.4: **State-space models and target tracking.** *Top Left:* Switching state-space model over discrete z_t and continuous x_t states, with observations y_t. *Bottom Left:* Dropping temporal correspondence on discrete states yields the target tracking model, where discrete components represent the *assignment* of observations to target or *clutter*. *Right:* Example output of target tracking with target observations (red) and clutter (blue).

showing improvements over expectation propagation.

1.3 Improved Variational Inference for State-Space Models

State-space models provide a framework for relating the unknown parameters of a time-evolving system (or *time-series* [154]) to a sequence of measurements from a, possibly noisy, remote sensor. These models are widely used in the acoustic arena for recognizing human speech [132] and distinguishing individual speakers [173], or to detect abnormal seismic events such as nuclear tests and earthquakes [103]. With more general sensors one is often interested in estimating an object's spatial position over time, known as *target tracking*.

Applications of target tracking are numerous, ranging from surveillance systems and air traffic control to visual object tracking. When the target in question is non-cooperating and can undergo changes in dynamics, these *maneuvers* must be inferred. A *switching state-space model* [14, 13] extends this representation to the case where the evolution of a continuous process is conditioned on a set of discrete states, Fig. 1.4 (left-top). To further complicate the problem advances in sensor technology often require tracking software to deal with high-sensitivity and low SNR environments. This results in another sort of discrete uncertainty in establishing the

correspondence between observations and targets, so-called *assignment uncertainty*. Such uncertainty is often modeled by dropping the temporal link between discrete elements, Fig. 1.4 (left-bottom).

In a probabilistic setting the assignment problem leads to computational intractability [5], for which there is a rich literature on approximation algorithms [12]. Stochastic approximations are widely used in the form of particle filters [45], but can be unstable for high-dimensional problems. Heuristic adaptations of the Kalman filter result in deterministic approximations with *nearest-neighbor data association* [5], which generally perform poorly in high clutter environments.

In Ch. 5 we extend the focus on deterministic methods and develop a family of algorithms based on (EP). This approach is similar in spirit to existing filtering approaches, but extends and generalizes these methods to produce smoothed posterior estimates. When compared to traditional tracking techniques on a variety of synthetic examples our approach produces significantly more accurate estimates.

1.4 Overview of Contributions

In this thesis we develop robust statistical methods for marginal and MAP inference, which we demonstrate on a variety of applications. We summarize our primary contributions below.

1.4.1 Particle-Based Continuous MAP Inference

In many domains involving models of complex–often physical–interactions it is necessary to estimate continuous marginals for which exact message updates are intractable. Moreover, increased dimensionality prohibits accurate numerical methods based on discretization. Monte Carlo methods like simulated annealing provide one common alternative [63, 6], but in many applications they are impractically slow to converge.

Inspired by work on *particle filters* and *sequential Monte Carlo* methods [30], several algorithms employ particle-based approximations of continuous BP messages via a non-uniform discretization which adapts and evolves across many message-passing iterations [92, 157, 82, 79]. This literature focuses on the sum-product BP algorithm for computing marginal distributions where importance sampling methods are used to update particle locations and weights.

Motivated by complementary families of MAP inference problems, we instead develop a *diverse particle max-product* (D-PMP) algorithm in Chapter 3. We view the

problem of approximating continuous max-product BP messages from an optimization perspective, and treat each particle as a hypothesized solution. Particle sets are kept to a computationally tractable size not by stochastic resampling, but by an optimization algorithm which directly minimizes errors in the max-product messages. We show that the D-PMP algorithm implicitly seeks to maintain all significant posterior modes, and is substantially more robust to initialization than previous particle max-product methods. We demonstrate the generality of this approach on a range of diverse applications such as human pose estimation, tracking, and protein structure prediction.

1.4.2 Continuous Optimization of Protein Side Chains

The high-dimensional representations used in protein structure prediction prohibit continuous minimization of the free energy. To cope with this most methods rely on a coarse discretization based on experimentally validated conformations [47]. Optimization proceeds either by a discrete surrogate objective or via simulated annealing with local gradient optimization [25, 29, 137]. While discrete approximations offer efficient alternatives, they fail to capture fine details of protein structure. Optimization based on simulated annealing can be computationally prohibitive and sensitive to initialization.

In Chapter 4 we apply particle max-product to protein side chain prediction, thereby avoiding the limitations associated with approximate techniques based on discretization. Moreover, our diverse particle selection procedure capably preserves multiple distinct configurations, thus avoiding local convergence issues which plague simulated annealing. Preserving these diverse conformations is important for structure prediction, where proteins assume multiple stable conformations relating to distinct functions [165, 100, 109].

1.4.3 Convergent Variational Inference without Degeneracy

Prior work in convergent inference is largely focused on discrete models, where degeneracy issues do not arise. For example, the belief optimization approach by Welling and Teh [172] exhibits convergence and degeneracy issues similar to loopy BP when applied to Gaussian models. The convex concave procedure [184] and related *double-loop* algorithms minimize the convex components of the variational objective while forming local convex approximations to the remaining terms. The double-loop algorithm of Heskes and Zoeter [73] is convergent for continuous switching state-space models, but does not address degeneracy.

In Ch. 5 we develop optimization of the variational objective which addresses the convergence and degeneracy issues that plague existing methods. Local convergence is guaranteed under mild assumptions via properties of the method of multipliers [22, 21]. Degenerate marginal approximations occur due to normalization constraints, which are inactive for discrete models, and that are not strictly enforced for continuous models. By using gradient projection methods we explicitly enforce these constraints and ensure iterates represent valid distributions at all stages of inference.

1.4.4 Improved Variational Inference for Tracking in Clutter

Probabilistic target tracking in the presence of missed and false (*clutter*) detections poses a challenging problem, for which exact Bayesian inference is intractable [5]. While there is thus a rich literature on approximate tracking algorithms we focus on deterministic approximate inference algorithms. Sequential Monte Carlo methods, such as particle filters, are also used for tracking [45] but lead to less compact state representations and can be unstable for high-dimensional problems.

The *probabilistic data association filter* (PDAF) [11] incorporates observations sequentially via a single forward pass, approximating the state's marginal distribution as Gaussian with matched mean and covariance. The *probabilistic multi-hypothesis tracker* (PMHT) [155, 11] instead adapts the *expectation maximization* (EM) algorithm to iteratively estimate smoothed state estimates from a fixed batch of data. These algorithms are derived from different measurement models: the PDAF assumes the target produces at most one true detection per time step, while the PMHT assumes the number of true detections is binomially distributed.

In Chapter 5 we propose a family of alternative tracking algorithms based on *expectation propagation* (EP) [114], a sophisticated variational approach to approximate inference. This approach is similar in spirit to the PDAF, in that we incorporate local evidence and project to a family of tractable approximate marginal distributions. Unlike PDAF, however, our EP algorithms can produce accurate smoothed state estimates; be easily adapted to various measurement models; and employ marginal approximations of varying complexity.

Chapter 2

Variational Inference for Graphical Models

We briefly introduce the concepts upon which our later contributions are based. We begin with a review of *graphical models* (Sec. 2.1), a core modeling tool which motivates our view of the statistical inference tasks we will explore in later chapters. From graphical models we move to the *exponential family* of distributions (Sec. 2.2). Aside from covering typical properties of the exponential family we introduce the *unnormalized exponential family*, which we will use in later sections related to expectation propagation (EP).

From the Markov independence properties of graphical models we construct efficient inference algorithms based on local computations (Sec. 2.3). These local computations can be interpreted as passing *messages* in the graphical model. *Message passing* algorithms optimize a global variational objective to perform approximate inference and are at the core of this thesis.

2.1 Graphical Models

Exploiting independence between random quantities is key to developing efficient algorithms for learning and inference, but modeling these interactions can be difficult. Graphical models offer a diagrammatic approach to modeling complex global relationships via simpler local interactions. Such representations are particularly helpful in complex systems, involving many random variables, where dependencies may be difficult to capture [127, 101, 37]. Moreover, the dependency structure encoded in the graphical model can be exploited to develop efficient inference algorithms.

Consider a set of random variables $x = (x_1, \ldots, x_d)$ and a collection of subsets $\mathcal{C}$ of indices $\{1, \ldots, d\}$ where each index appears in at least one subset. A probability

density function (PDF) can be represented as the following product density,

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c), \qquad Z = \int_{\mathcal{X}} \prod_{c \in \mathcal{C}} \psi_c(x_c) dx_c. \tag{2.1}$$

The quantity Z is a global normalizing constant which is independent of random variables x. Through connections to statistical physics Z is called the *partition function*. With a slight abuse of terminology we refer to the nonnegative factors ψ as *potential functions*. The connection with energy potentials is more explicitly shown by the equivalent *Gibbs distribution*,

$$p(x) \propto \exp\left(-\frac{1}{T} \sum_{c \in \mathcal{C}} \varphi_c(x_c) \right), \tag{2.2}$$

where $\varphi_c = -\log \psi_c$ are the energy potentials. Traditionally the Gibbs representation explicitly includes a *temperature* T, which controls the relative height of modes in the distribution. Every random vector x has a Gibbs representation, though the factorization is not necessarily unique. For example, consider the two equivalent Gibbs representations,

$$\psi_{123}(x_1, x_2, x_3)\psi_4(x_4) \quad \text{and} \quad \psi_{12}(x_1, x_2)\psi_{34}(x_3, x_4). \tag{2.3}$$

While unique these factorizations may express the same probability density, and the dependency structure of each factorization may yield inference algorithms with different computational properties.

Graphical models can be partitioned into two classes based on edge type: *undirected* and *directed*. Undirected models are more general and often easier to specify than their directed counterparts because factors underlying the graph need not be locally normalized. However, directed models often result in more efficient inference algorithms and can simplify the sampling process due to the restriction that factors are conditional probabilities. In the following sections we discuss three of the most common graphical model types, depicted in Figure 2.1, along with the benefits and limitations of each.

2.1.1 Undirected Graphical Models

The graph $G = (\mathcal{V}, \mathcal{E})$ defines a graphical model with vertices $s \in \mathcal{V}$ and edges $(s, t) \in \mathcal{E}$. Undirected models have the property that edges are defined irrespective of node ordering,

$$(s, t) \in \mathcal{E} \iff (t, s) \in \mathcal{E}. \tag{2.4}$$

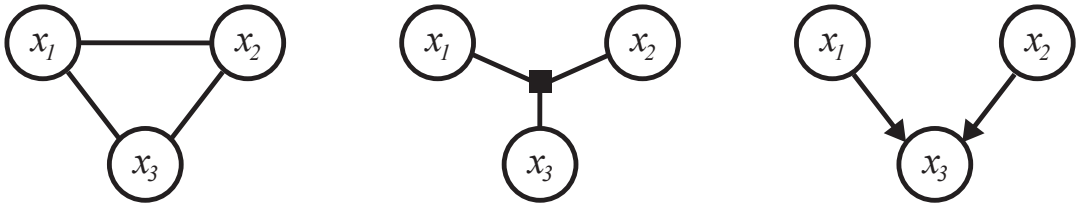


Figure 2.1: **Graphical models.** *Left:* A pairwise Markov random field respects the factorization $\psi(x_1, x_2)\psi(x_2, x_3)\psi(x_1, x_3)$. The factorization is implicit and may equally respect a non-pairwise factorization $\psi(x_1, x_2, x_3)$. *Center:* A factor graph explicitly denotes the latter non-pairwise factorization. *Right:* A Bayesian network encodes an unambiguous product of normalized conditional probabilities $p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$.

The most general construction of an undirected graphical model is given by the *Markov random field* (MRF). The MRF implicitly encodes the Gibbs factorization (2.1) in terms of graph cliques. The *factor graph* extends this construction by explicitly denoting the factorization via special factor nodes. The following sections describe each construction in detail.

Markov Random Fields

The MRF dates back to a study of random lattices by Dobruschin [44] and has since been widely used in applications ranging from statistical physics [181] to low-level computer vision [63, 71]. An MRF is an undirected graph $G = (\mathcal{V}, \mathcal{E})$ with a vertex $s \in \mathcal{V}$ for each component variable x_s. An edge $(s, t) \in \mathcal{E}$ exists for every pair of nodes s and t that appear in a potential $\psi_c(\cdot)$ of joint factorization,

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c). \tag{2.5}$$

A stochastic process x is said to be *Markov with respect to the graph* $G = (\mathcal{V}, \mathcal{E})$ if x has a Gibbs representation that gives the graph G by the above construction.

An MRF is equivalently defined in terms of conditional independencies. For a disjoint subset of nodes A, B and C let,

$$x_A \perp x_C \mid x_B \tag{2.6}$$

mean the variables x_A are independent of x_C conditioned on x_B. A stochastic process x is Markov with respect to G if for every set of conditionally independent variables $x_A \perp x_C \mid x_B$ any path from vertex A to C must pass through vertex B. If x is Markov with respect to G then x is Markov with respect to any graph $G' = (\mathcal{V}, \mathcal{E}')$

over the same vertices $\mathcal{V}$ that has $\mathcal{E} \subseteq \mathcal{E}'$, and by this reasoning x always respects the complete graph.

The factorization expressed by an MRF G is implied by the graph cliques. For each potential $\psi_c(x_c)$ there must exist a clique in G with vertices $c \subseteq \mathcal{V}$. Moreover, the factorization expressed by an MRF may not be unique, for example the complete graph is consistent with any factorization.

A *minimal* representation is one in which the factors ψ_c are maximal cliques of the MRF G. But even this minimal representation may not be unique. For example, consider the function which takes the value one if $x_1 = x_2 = x_3$ and zero everywhere else. There are two minimal factorizations of this function,

$$\mathbb{I}(x_1 = x_2)\mathbb{I}(x_1 = x_3) \quad \text{and} \quad \mathbb{I}(x_1 = x_2)\mathbb{I}(x_2 = x_3), \tag{2.7}$$

where $\mathbb{I}$ is the Kronecker delta function.

Pairwise Markov Random Fields

Frequently throughout this thesis we will express $p(x)$ as a product density which factorizes according to single-node *local evidence* potentials, and pairwise *compatibility* potentials. This construction is known as a *pairwise MRF* and takes the form,

$$p(x) = \frac{1}{Z} \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t). \tag{2.8}$$

Pairwise MRFs are simple to specify which has lead to their wide adoption, but they are also flexible. Pairwise MRFs are a general class of graphical models in the sense that they can encode any probability distribution via its junction tree representation [41, 37]. Specialized classes of MRFs such as the pairwise binary lattice date back to the pioneering work of Ising [84], the so-called *Ising model*, as well as the multivariate extension, the Potts model. For these reasons we will assume pairwise MRF models in many of the later chapters.

Factor Graphs

Factor graphs explicitly represent the Gibbs factorization through special *factor nodes*, and thus avoid the ambiguities associated with MRFs. The representation is an undirected bipartite graph $G = (\mathcal{V}, \mathcal{E})$ with a vertex for every variable x_s and each factor ψ_c. An edge $(s, c) \in \mathcal{E}$ connects these two nodes if x_s is an argument of ψ_c. For $\mathcal{C} \subset \mathcal{V}$, the set of factor vertices, the joint density of a factor graph is unique,

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(x_c), \tag{2.9}$$

where x_c is the set of variables connected by the factor vertex $c \in \mathcal{C}$. Each edge connects a random variable with a factor as depicted in Fig. 2.1.

The topology of factor graphs enables development of efficient *message passing* inference such as the sum-product belief propagation algorithm [96], which we discus in Sec. 2.3.2. The connection between graphical representations and inference has made factor graphs popular for developing channel decoders such as the popular turbo codes [20, 19], shown to be equivalent to sum-product BP inference on a factor graph [56, 112]. The factor graph representation was then used to extend turbo codes and rediscover the class of low density parity check (LDPC) codes [60, 32, 20, 19, 33].

2.1.2 Bayesian Networks

When factors are locally normalized conditional probabilities we may define a directed acyclic graph (DAG), known as a Bayesian network. The graph $G = (\mathcal{V}, \mathcal{E})$ contains a single node for each random variable $s \in \mathcal{V}$ with directed edges, so the existence of a directed edge $(s,t) \in \mathcal{E}$ does not imply the reverse edge exists,

$$(s,t) \in \mathcal{E} \nRightarrow (t,s) \in \mathcal{E}. \tag{2.10}$$

For the distribution to be well-defined the graph must not contain directed cycles. More formally, for any directed sequence $(s_1, \ldots, s_k)$ where $(s_i, s_{i+1}) \in \mathcal{E}$, there must be no edge from k back to the start node $(s_k, s_1) \notin \mathcal{E}$.

For a node $s \in \mathcal{V}$ let $\mathrm{Pa}(s)$ be the set of *parents*. Each factor is a normalized conditional probability,

$$\psi(x_s, x_{\mathrm{Pa}(s)}) = p(x_s \mid x_{\mathrm{Pa}(s)}). \tag{2.11}$$

We can write the full joint distribution as a product of conditional and marginal probabilities. For a leaf node without parents we follow the convention that $\mathrm{Pa}(s) = \emptyset$ and the factor is a prior probability $\psi(x_s, x_{\mathrm{Pa}(s)} = \emptyset) = p(x_s)$:

$$p(x) = \prod_{s \in \mathcal{V}} p(x_s \mid x_{\mathrm{Pa}(s)}). \tag{2.12}$$

Bayesian networks simplify marginal inference since we do not need to compute a partition function. Sampling the distribution is also more straightforward than for undirected graphs. To draw a sample $x \sim p(x)$ one begins by drawing samples from the prior distribution at the leaves $x_s \sim p(x_s)$. Then, one samples the *children* of node s, denoted $\mathrm{Ch}(s)$, so that $x_t \sim p(x_t \mid x_{\mathrm{Pa}(t)})$. This *generative process* is repeated until all nodes are sampled and it is guaranteed to be well-defined on a DAG.

2.2 Exponential Family

The *exponential family* encompasses a large class of well-known probability distributions including the Gaussian, Bernoulli, Multinomial, Gamma, Dirichlet, and many others. The exponential family is well-studied both in terms of its analytic properties [169] and its information geometry [135, 39, 4, 3, 16], making it a convenient set to work with. This is particularly true for statistical inference tasks, which can often be recast as mapping between alternate forms of parameterization. In the following sections is a brief overview of some of the properties directly relevant to later chapters, beginning with a definition of the exponential family and its alternate forms of parameterization.

2.2.1 Definition and Parameterization

The exponential family defines a distribution for a random vector $x \in \mathcal{X}$ in terms of *canonical parameters* $\theta \in \Theta$ and *sufficient statistics* $\phi(x) = (\phi_1(x), \ldots, \phi_d(x))$ where $\phi_\alpha : \mathcal{X} \to \mathbb{R}^d$. The corresponding density is fully specified by these elements,

$$p_\theta(x) = h(x) \exp\left\{\theta^T \phi(x) - \Phi(\theta)\right\}, \qquad \Phi(\theta) = \log \int_{\mathcal{X}} h(x) \exp\left\{\theta^T \phi(x)\right\} dx, \quad (2.13)$$

with *base measure* $h(x)$ and log-partition function $\Phi(\theta) = \log Z(\theta)$. The log-partition function plays a prominent role in later chapters, and estimating it is a core problem in statistical inference where it is the marginal log-likelihood of the data. In exponential families the log-partition function is also the *cumulant generating function*. By this relationship derivatives of the partition function yield moments of the distribution:

$$\frac{\partial \Phi(\theta)}{\partial \theta_\alpha} = \mathbb{E}_{p_\theta}[\phi_\alpha(x)], \qquad \frac{\partial^2 \Phi(\theta)}{\partial \theta_\alpha \partial \theta_\beta} = \mathbb{E}_{p_\theta}[\phi_\alpha(x)\phi_\beta(x)]. \quad (2.14)$$

That the second derivative yields a covariance, which by definition is positive semi-definite, implies convexity of $\Phi(\theta)$ [169, Prop. 3.1]. This fact implies convexity on the set of valid canonical parameters,

$$\Theta \triangleq \{\theta \in \mathbb{R}^d \mid \Phi(\theta) < +\infty\}. \quad (2.15)$$

An alternative parameterization can be given in terms of *mean parameters* $\mu_\alpha = \mathbb{E}_{p_\theta}[\phi_\alpha(x)]$. We will generally use p_μ to denote the mean parameterization and p_θ for the canonical parameterization. Convexity of $\Phi(\theta)$, along with the cumulant generating function property (2.14), implies convexity of the set of *realizable mean parameters*,

$$\mathcal{M} \triangleq \{\mu \in \mathbb{R}^d \mid \exists\, p \text{ s.t. } \mathbb{E}_\theta[\phi(x) = \mu]\}. \quad (2.16)$$

Characterizing this set plays a prominent role in variational inference (Sec. 2.3) as it provides a constraint set that defines valid distributions. In Sec. 5.2 we develop marginal inference based on gradient projection, which avoids degenerate distributions outside of this set, such as Gaussians with negative definite covariance matrices.

2.2.2 Basic Properties

The *canonical form* of (2.13) has a number of useful properties, in particular exponential families are closed under multiplication and the product density takes a simple form. For density functions $f(x)$ and $q(x)$ with canonical parameters θ_f and θ_q, respectively, the product density $p(x) \propto f(x)q(x)$ is in the exponential family with parameters $\theta_p = \theta_f + \theta_q$ provided $\theta_p \in \Theta$ are valid canonical parameters,

$$\exp\left\{\theta_f^T\phi(x) - \Phi(\theta_f)\right\}\exp\left\{\theta_q^T\phi(x) - \Phi(\theta_q)\right\} \propto \exp\left\{(\theta_f + \theta_q)^T\phi(x) - \Phi(\theta_f + \theta_q)\right\}.$$

A corollary of this is the well-known property that Gaussians are closed under multiplication. A less well-known property that we will exploit in later sections, is that Gaussians are closed under division, provided the resultant covariance matrix is positive semidefinite.

Canonical parameters are useful for analytic operations, such as multiplication and division, whereas mean parameters are more useful for statistical tasks. For instance, given some $\widetilde{p}(x)$ we can find the closest $p_\mu(x)$ in the exponential family in terms of Kullback-Leibler divergence. This *I-projection*, as it is known in information geometry, leads to the well-known *moment-matching* property of the exponential family,

$$\hat{\mu} \triangleq \arg\min_{\mu} \; KL(\widetilde{p}\|p_\mu) \Leftrightarrow \hat{\mu} = \mathbb{E}_{\widetilde{p}}[\phi(x)]. \tag{2.17}$$

This property states that for any distribution $\widetilde{p}(x)$ the closest exponential family approximation is efficiently found by calculating the expected sufficient statistics $\mathbb{E}_{\widetilde{p}}[\phi(x)]$.

Some exponential families are closed under marginalization, for example consider the multivariate Gaussian $N(x \mid \mu, \Sigma)$. The marginal distribution $p(x(i))$, the i^{th} element of x, is given by the corresponding elements of the mean parameters $N(x(i) \mid \mu(i), \Sigma(i,i))$. An example of an exponential family not closed under marginalization is given by the product of Gaussian and multinomial distributions $N(x \mid \mu_k, \Sigma_k)Mult(k \mid \theta)$. Marginalization over k yields a Gaussian mixture which is not in the exponential family.

2.3 Message Passing and Variational Inference

The approach underlying variational inference is to recast statistical inference as an optimization problem. The resulting objective, known as the *variational free energy*, minimizes Kullback-Leibler divergence with respect to an unknown function of the random variables (Sec. 2.3.1). Variational inference derives its name from the calculus of variations, which addresses the problem of optimizing functionals [62]. In practice, the optimization is typically intractable, both in the number of constraints and terms in the objective function, so approximations and relaxations are introduced. Fixed-point iterations known as *message passing algorithms* approximate optimization for marginal (Sec. 2.3.2) and *maximum a posteriori* (MAP) inference (Sec. 2.3.3). A good tutorial on the subject of variational inference is provided by Jaakkola [85] with further details in [88, 169].

2.3.1 Variational Free Energy

We begin by introducing the concept of marginal inference and the corresponding variational formulation. Given a distribution $p(x, y)$ let y be *observed* values and let x be *latent* random vectors. Consider a variable x_s, the task of *posterior marginal inference* is to compute the conditional distribution,

$$p(x_s \mid y) = \frac{p(x_s, y)}{p(y)}, \quad p(y) = \int p(x, y) dx. \tag{2.18}$$

More generally, we may wish to compute the conditional density $p(x_S \mid y)$ for any subset of variables $x_S = (x_{s_1}, \ldots, x_{s_n})$. The normalization $p(y)$ is the *marginal likelihood* – it depends on observed data and any model parameters.

Variational inference poses the following optimization,

$$\underset{q}{\text{maximize}}\ J(q) = \log p(y) - \mathrm{KL}(q(x) \,\|\, p(x \mid y)). \tag{2.19}$$

Maximization is with respect to the *variational distribution* q, which is constrained to be a valid probability distribution. While the optimization (2.19) seems vacuous, it is sensible for a couple of reasons: first, observe that the Kullback-Leibler divergence is non-negative and vanishes when $q(x)$ equals the true posterior $p(x \mid y)$. Second, if the true posterior is a feasible solution then the optimal value $J(q^*)$ equals the log-partition function, and for all other values of q yields a bound,

$$J(q) \leq \log p(y), \quad J(q^*) = \log p(y).$$

However, the variational optimization (2.19) can only be evaluated if the posterior $p(x \mid y)$ is tractable. With some algebra, the objective can be reformulated in a way

that can be evaluated for any tractable distribution q,

$$J(q) = -\mathrm{KL}(q(x) \,\|\, p(x,y)) = \mathbb{E}_q[-\log p(x,y)] + \mathcal{H}(q). \tag{2.20}$$

In this form the objective decomposes according to the structure of q and the joint distribution p. Moreover, the variational objective (2.20) is closely related to energy minimization in statistical physics. Let $\mathcal{F}(q) = -J(q)$ and we have the *variational free energy*,

$$\underset{q}{\text{minimize}} \;\; \mathcal{F}(q) = \mathbb{E}_q[-\log p(x,y)] - \mathcal{H}(q). \tag{2.21}$$

The two terms on the r.h.s. of the variational free energy (2.21) have competing influence. The first term $\mathbb{E}_q[\cdot]$ is known as the *average energy* and encourages the variational distribution q to explain the data. The second term $\mathcal{H}(q)$ acts as a regularizer by encouraging maximum entropy.

2.3.2 Message Passing for Marginal Inference

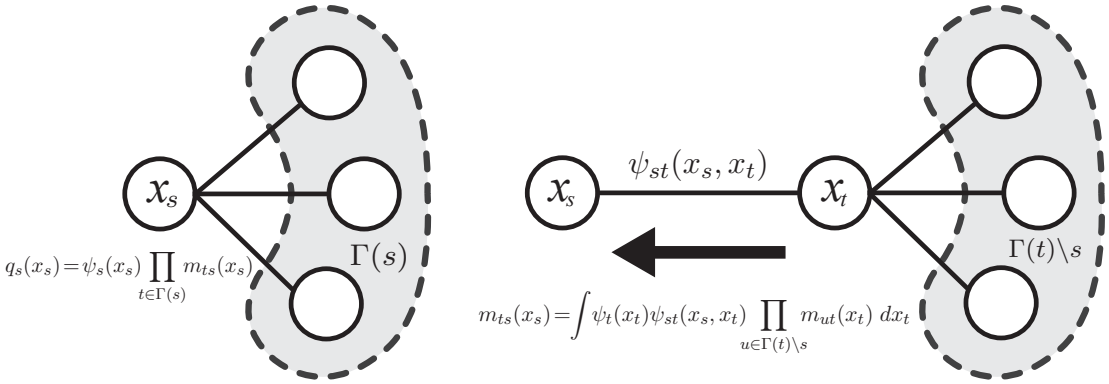
For tree-structured distributions the variational problem (2.21) can be efficiently optimized via fixed point algorithms whereby each iteration recursively updates local statistics in the graphical model. These fixed point iterations can be interpreted as passing messages along edges of the graph. The resulting algorithms, known as *message passing* algorithms, decompose the global variational objective, via Markov independence, and optimize the free energy in terms of strictly local computations. For graphs with cycles we will see that these algorithms are not guaranteed to yield exact inference, but that empirically they often produce accurate approximations.

Belief Propagation

We show how Markov independence leads to efficient inference through a concrete example. To simplify the discussion let us drop the distinction between *observed* data and focus only on latent variables. Consider the following pairwise MRF,

$$p(x_1,x_2,x_3,x_4) \propto \psi_1(x_1)\psi_2(x_2)\psi_3(x_3)\psi_4(x_4)\psi_{12}(x_1,x_2)\psi_{23}(x_2,x_3)\psi_{24}(x_2,x_4). \tag{2.22}$$

For an acyclic model the marginal $p(x_1)$ can be computed directly, and we can use properties of Markov independence to decompose this computation. In this example x_2 is a *separator* for any pair of variables, meaning any two variables are conditionally independent given x_2 (see Sec. 2.1.1). Therefore, we can compute integrals over x_3 and x_4 independently and multiply the results to integrate over x_2. We denote these



Belief **Message Update**

Figure 2.2: Belief propagation update equations for marginal and message.

sub-operations $m_{32}(x_2)$ and $m_{42}(x_2)$ because they can be interpreted as *messages* being passed between nodes in the graphical model:

$$p(x_1) \propto \psi_1(x_1) \underbrace{\int \psi_2(x_2) \underbrace{\left(\int \psi_{23}(x_2, x_3)\, dx_3 \right)}_{m_{32}(x_2)} \underbrace{\left(\int \psi_{24}(x_2, x_4)\, dx_4 \right)}_{m_{42}(x_2)} dx_2}_{m_{21}(x_1)} . \tag{2.23}$$

Belief Propagation (BP) [127] codifies the marginal calculations for distributions Markov with respect to a graph G. The BP marginal and message update equations for a pairwise MRF are,

$$q(x_s) \propto \psi_s(x_s) \prod_{t \in \Gamma(s)} m_{ts}(x_s) \tag{2.24}$$

$$m_{ts}(x_s) = \int \psi_t(x_t)\psi_{st}(x_s, x_t) \prod_{u \in \Gamma(t)\setminus s} m_{ut}(x_t)\, dx_t \tag{2.25}$$

The marginal over x_s is given by the product of messages $m_{ts}(x_s)$ from neighbors $t \in \Gamma(s)$, and the local evidence $\psi_s(x_s)$. The message from node t to s is computed recursively by multiplying incoming messages to node x_t with the local evidence and compatibility potentials, and then integrating over x_t. The message is a function of x_s, the receiving node. Figure 2.2 gives a graphical representation of the message and marginal updates.

The pairwise factorization is assumed for simplicity since any model can be expressed as a pairwise MRF via its junction tree representation (c.f. Sec. 2.1.1). For general factor graphs BP is variously known as the *sum-product* BP algorithm [96], since updates involve products over messages, and for discrete models the integrals become summations.

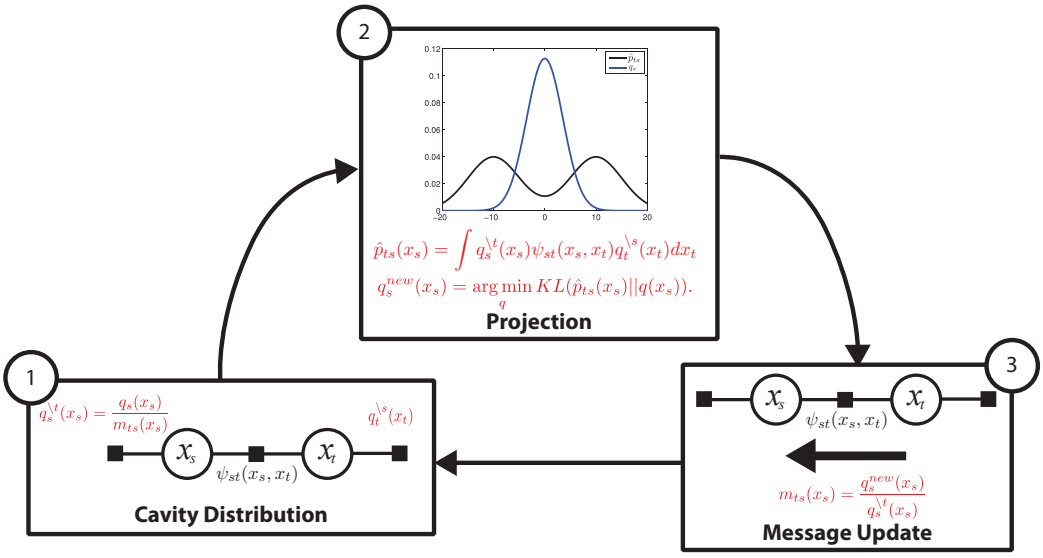


Figure 2.3: Expectation propagation updates for pairwise MRFs.

For acyclic graphs BP yields exact marginals. For graphs with cycles the recursions (2.25) are well-defined, but they produce approximations to the true marginals. *Loopy BP* (LBP) often produces good approximations in practice [119, 112], but it is not guaranteed to converge. Loopy BP convergence is well-studied for discrete pairwise MRFs resulting in sufficient conditions for convergence. Early work by Tatikonda and Jordan [161] drew on connections with a theory of Gibbs measures to insure convergence of LBP based on Dobrushin's condition. These results were later tightened by Heskes [72], and further by Ihler [78] with analysis that implies bounds on the distance between LBP fixed points and on the propagation of message errors.

Expectation Propagation

The message update integral (2.25) constrains BP to discrete and Gaussian MRFs. Expectation Propagation (EP) applies to a much broader class of models [114, 73, 169], and is made possible by exploiting closure properties of the exponential family (see Sec. 2.2). For a cleaner presentation we assume a pairwise MRF consisting of only pairwise potentials,

$$p(x) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t). \tag{2.26}$$

This factorization is w.l.o.g. since node factors ψ_s can always be absorbed arbitrarily into a neighboring edge potential. Expectation propagation approximates the marginal $p_s(x_s)$ with a density in the exponential family defined to be a product of

messages,

$$q_s(x_s) = \exp(\langle \theta_s, \phi(x_s) \rangle - \Phi(\theta_s)) \propto \prod_{t \in \Gamma(s)} m_{ts}(x_s) \tag{2.27}$$

with canonical parameters θ_s, sufficient statistics $\phi(x_s)$ and mean parameters $\mu_s = \mathbb{E}[\phi(x_s)]$. The messages $m_{ts}(\cdot)$ belong to the *unnormalized* exponential family with parameters θ_{ts} and a scale factor γ_{ts},

$$m_{ts}(x_s) = \gamma_{ts} \exp(\langle \theta_{ts}, \phi(x_s) \rangle). \tag{2.28}$$

To update the marginal approximation $q_s(x_s)$ we first choose a factor ψ_{st} and remove the corresponding messages from marginals over nodes incident to this factor. The so-called *cavity* is an unnormalized exponential family:

$$q_s^{\backslash t}(x_s) = \frac{q_s(x_s)}{m_{ts}(x_s)} = \gamma_s^{\backslash t} \exp(\langle \theta_s^{\backslash t}, \phi(x_s) \rangle). \tag{2.29}$$

We form the *augmented distribution* by multiplying the true factor ψ_{st} with the corresponding cavity distributions and integrating over x_t,

$$\hat{p}_{ts}(x_s) = \int q_s^{\backslash t}(x_s) \psi_{st}(x_s, x_t) q_t^{\backslash s}(x_t) dx_t. \tag{2.30}$$

The augmented distribution (2.30) is a local approximation to the marginal, but is not necessarily in the exponential family. The variational approximation is updated by projecting into the exponential family,

$$q_s^{new}(x_s) = \arg\min_q \ \ \mathrm{KL}(\hat{p}_{ts}(x_s) \,\|\, q(x_s)). \tag{2.31}$$

Using the moment matching property of the exponential family (see Sec. 2.17) we update the parameters as,

$$\mu_s^{new} = \mathbb{E}_{q_s^{new}}[\phi(x_s)] = \mathbb{E}_{\hat{p}_{ts}}[\phi(x_s)]. \tag{2.32}$$

One restriction EP imposes is that moments of the augmented distribution (2.32) can be computed. If these integrals are not analytic they can be numerically approximated, for example by quadrature methods [186]. The associated canonical parameters θ_s^{new} can be computed from (2.32) to yield the log-partition $\Phi(\theta_s^{new})$, which fully specifies the exponential family density. Finally, we update the message from t to s as,

$$m_{ts}(x_s) = \frac{q_s^{new}(x_s)}{q_s^{\backslash t}(x_s)}. \tag{2.33}$$

This message update is easily calculated by subtracting canonical parameters, $\theta_{ts} = \theta_s^{new} - \theta_s^{\backslash t}$. The algorithm proceeds iteratively by updating each factor ψ_{st} for all $(s,t) \in \mathcal{E}$ in any order. Convergence is determined when the change in message parameters θ_{ts} falls below some specified threshold.

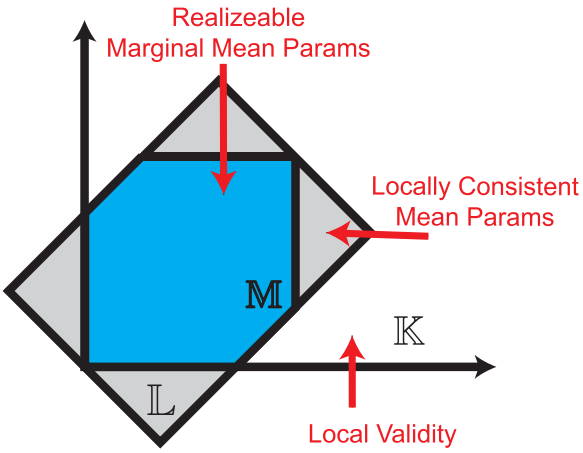


Figure 2.4: Constraint sets of Bethe variational problem (2.39)

Bethe Variational Objective

Fixed points of BP and EP correspond to stationary points of a variational optimization known as the *Bethe free energy* [183, 181, 182]. We briefly discuss this correspondence for a pairwise MRF $G = (\mathcal{V}, \mathcal{E})$, a more detailed discussion is given by Yedidia et al. [183]. Consider a variational distribution $q(x)$ in the exponential family with sufficient statistics $\phi(x) \in \mathbb{R}^d$,

$$q(x) \propto \exp\{\langle \theta, \phi(x) \rangle\}, \quad \mu = \mathbb{E}_q[\phi(x)]. \tag{2.34}$$

The variational distribution decomposes into unary $q_s(x_s)$ and pairwise $q_{st}(x_s, x_t)$ marginal approximations. With this parameterization we obtain a compact representation of the variational free energy (2.21) in terms of mean parameters,

$$\min_{\mu \in \mathbb{M}(G)} \mathcal{F}(\mu) = \min_{\mu \in \mathbb{M}(G)} \mathbb{E}_\mu[-\log p(x)] - \mathcal{H}[\mu]. \tag{2.35}$$

The constraint set $\mathbb{M}(G)$ is the set of *realizable marginal mean parameters*, which ensures that the variational distributions are well-defined marginals,

$$\mathbb{M}(G) = \{\, \mu : \exists \text{ some } p(x) \text{ with marginal mean parameters } \mu \,\}. \tag{2.36}$$

For discrete models $\mathbb{M}(G)$ is known as the *marginal polytope* and is specified by a set of linear inequalities. However, exactly characterizing the marginal polytope may require exponentially many constraints [169]. We relax the constraints to the set of *locally consistent* marginal distributions $\mathbb{L}(G)$, which are properly normalized and satisfy expectation constraints associated with each edge of the graph,

$$\underbrace{C_s(\mu) = 1 - \int q_s(x_s; \mu_s)\, dx_s}_{\text{Normalization}}, \quad \underbrace{C_{ts}(\mu) = \mu_s - \mathbb{E}_{q_{st}}[\phi_s(x_s)]}_{\text{Local Consistency}}. \tag{2.37}$$

This is a relaxation in the sense that $\mathbb{M}(G) \subset \mathbb{L}(G)$ with strict equality if G does not contain cycles. We approximate the entropy $\mathcal{H}[\mu]$ with the entropy of a tree-structured distribution. Such an approximation is tractable and consistent with $\mathbb{L}(G)$, and yields the *Bethe free energy*,

$$\begin{aligned}\mathcal{F}_B(\mu) = &\sum_{(s,t)\in\mathcal{E}} \mathbb{E}_{q_{st}}[\log q_{st}(x_s, x_t) - \log \varphi_{st}(x_s, x_t)] \\ &- \sum_{s\in\mathcal{V}} (n_s - 1)\, \mathbb{E}_{q_s}[\log q_s(x_s) - \log \psi_s(x_s)],\end{aligned} \tag{2.38}$$

where we define the shorthand $\varphi_{st} = \psi_{st}\psi_s\psi_t$ and $n_s = |\Gamma(s)|$ is the number of neighbors to node s. The resulting *Bethe variational problem* (BVP) is,

$$\begin{aligned}\underset{\mu}{\text{minimize}} \quad & \mathcal{F}_B(\mu) \\ \text{subject to} \quad & C_{ts}(\mu) = 0, \forall s \in \mathcal{V}, t \in \Gamma(s) \\ & C_s(\mu) = 0, \forall s \in \mathcal{V}, \\ & \{\mu_s : s \in \mathcal{V}\} \cup \{\mu_{st} : (s,t) \in \mathcal{E}\} \in \mathbb{K}.\end{aligned} \tag{2.39}$$

The constraint set $\mathbb{K} = \bigcup_s \mathbb{K}_s \bigcup_{st} \mathbb{K}_{st}$ defines the set of valid mean parameters μ. The definition of $\mathbb{K}$ depends on the variational distribution q, for example if q is Gaussian then $\mathbb{K}$ is the positive semidefinite cone.

2.3.3 Message Passing for MAP Inference

Another common task in statistical inference is to quantify uncertainty about the maximizing configuration of random variables, known as *maximum a posteriori* (MAP) inference. In this section we show that the MAP problem has a variational formulation which can be seen as the *zero temperature limit* of the variational problem [169, 168, 166]. We also introduce the max-product (MP) variant of BP, a message passing algorithm which solves the MAP variational objective in tree-structured graphical models [2, 37, 41]. For loopy models, where the variational problem is intractable, we discuss the reweighted max-product (RMP) algorithm, which minimizes an upper bound on the MAP probability.

2.3.4 Variational MAP Inference

Maximum a posteriori inference is sensible in applications where a jointly consistent estimator is preferred. Some examples we discuss later in this thesis are the articulated models of human pose (Sec. 3.3.1 and Sec. 3.3.2) and protein structure prediction (Ch. 4).

To simplify the presentation we assume $p(x)$ is a discrete pairwise MRF with $x_s \in \{1, \ldots, K\}$. The log-joint distribution is given by the following *overcomplete representation*,

$$\log p(x) = \sum_{s \in \mathcal{V}} \sum_{i=1}^{K} \theta_{s;i} \delta_i(x_s) + \sum_{(s,t) \in \mathcal{E}} \sum_{i=1}^{K} \sum_{j=1}^{K} \theta_{st;ij} \delta_i(x_s) \delta_j(x_t) + \text{const.} \tag{2.40}$$

Let $\phi_s(x_s) = (\delta_1(x_s), \ldots, \delta_K(x_s))^T$ be a vector of sufficient statistics for node $s \in \mathcal{V}$. The set of maximizing configurations x^* is given by the more compact representation:

$$x^* \in \arg\max_x \; \log p(x) = \arg\max_x \; \langle \theta, \phi(x) \rangle. \tag{2.41}$$

The *max-marginal distribution* encodes uncertainty about the maximum value for any variable, and is the MAP analogue of the posterior marginal:

$$q_s(x_s) \propto \underset{x'}{\text{maximize}} \; p(x') \;\; \text{subject to } x'_s = x_s. \tag{2.42}$$

With the exponential family assumption we can formulate the variational problem (2.19) as a maximization w.r.t. mean parameters,

$$\Phi(\theta) = \underset{\mu}{\text{maximize}} \; \langle \theta, \mu \rangle + \mathcal{H}(\mu). \tag{2.43}$$

A connection between MAP and variational inference is drawn by considering an inverse scaling of the canonical parameters θ/T. This *temperature* T plays the same role as in the Gibbs distribution (2.2); it controls the relative height between modes. As $T \to 0$ the distribution places all of its mass on the set of maximizers (2.41). The MAP problem is recovered as the zero temperature limit of the variational problem [169, 168, 166],

$$\max_x \langle \theta, \phi(x) \rangle = \lim_{T \to 0} \; T\Phi\left(\frac{\theta}{T}\right) = \lim_{T \to 0} \; \max_\mu \; \{\langle \theta, \mu \rangle - T\,\mathcal{H}(\mu)\}\,. \tag{2.44}$$

Taking the limit $T \to 0$ in (2.44) yields the result that MAP inference corresponds to a linear program (LP),

$$\max_x \langle \theta, \phi(x) \rangle = \max_{\mu \in \mathbb{M}} \langle \theta, \mu \rangle \leq \max_{\mu \in \mathbb{L}} \langle \theta, \mu \rangle. \tag{2.45}$$

Derivation of the relaxation above is analogous to the development of the Bethe free energy discussed in Sec. 2.3.2. The constraint set (the marginal polytope $\mathbb{M}$) is intractable for arbitrary models making the optimization NP-hard [149]. The set of locally consistent marginals $\mathbb{M} \subset \mathbb{L}$ relaxes these constraints and produces a tractable approximation. The assumption that x is discrete ensures that the inner product $\langle \theta, \mu \rangle$ is well-defined, but a similar correspondence holds for continuous MRFs with a suitable inner product definition [170, 116, 128].

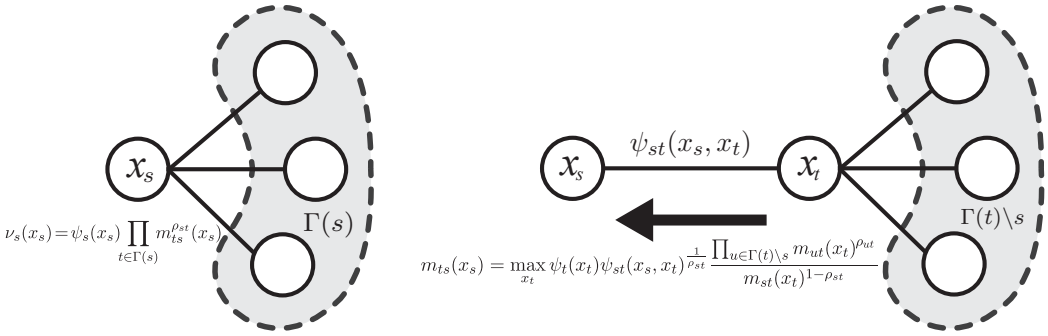


Figure 2.5: Reweighted max-product (RMP) updates.

Max-Product Belief Propagation

Just as sum-product BP optimizes the Bethe free energy, we seek a message passing algorithm that solves the MAP LP relaxation (2.45). The max-product (MP) variant of BP replaces integration with maximization to update messages,

$$m_{ts}(x_s) = \underset{x_t}{\text{maximize}}\ \psi_t(x_t)\psi_{st}(x_s, x_t) \prod_{u \in \Gamma(t) \backslash s} m_{ut}(x_t). \tag{2.46}$$

For tree-structured distributions MP solves the MAP LP relaxation (2.45), which is tight, and MP beliefs correspond to the true max-marginal distributions [168]. For models with cycles, however, the connection is more subtle. We adopt the terminology of Wainwright et al. [169] and refer to max-product beliefs as *psuedo-max-marginals*,

$$\nu_s(x_s) \propto \psi_s(x_s) \prod_{t \in \Gamma(s)} m_{ts}(x_s). \tag{2.47}$$

While MP does not solve the MAP LP relaxation for general models and in the next section we consider a reweighted variant of max-product more closely aligned with the MAP LP objective.

Upper Bounds on the MAP Probability

The MAP LP relaxation (2.45) allows us to verify global optimality when the solution to the MAP LP relaxation $\mu^* = \arg\max_{\mu \in \mathbb{L}} \langle \mu, \theta \rangle$ is integral. However, max-product does not solve the MAP LP relaxation for arbitrary models, except those with special combinatorial structure such as bipartite matching and weighted b-matching [169].

An alternative bound of the MAP probability is given by a convex combination of tree-structured MRFs. We begin with a distribution ρ over spanning trees $\mathcal{T}$ for some graph G:

$$\rho = \{\rho(T) \mid \rho(T) \geq 0, \sum_{T \in \mathcal{T}} \rho(T) = 1\}, \tag{2.48}$$

Associated with each tree $T \in \mathcal{T}$ is a distribution with canonical parameter vector $\theta(T)$ respecting the tree structure. Jensen's inequality upper bounds the MAP log-probability:

$$\underset{x}{\text{maximize}}\ \langle \theta, \phi(x) \rangle \leq \sum_{T \in \mathcal{T}} \rho(T)\ \underset{x}{\text{maximize}}\ \langle \theta(T), \phi(x) \rangle. \tag{2.49}$$

The bound (2.49) involves a maximization over tree-structured distributions, and so it can be evaluated efficiently for fixed ρ and parameters $\theta(T)$. The bound is tight if and only if all tree distributions in the support of ρ agree on their maximizers – a condition is known as *tree agreement*.

Given a fixed spanning tree distribution ρ the set of valid tree parameters $\theta(T)$ must equal parameters of the true distribution θ when averaged over the spanning tree distribution $\mathbb{E}_\rho[\theta(T)] = \theta$. The dual problem of finding the tightest bound (2.49) can be formulated by the following LP,

$$\underset{\theta(T)}{\text{minimize}} \sum_{T \in \mathcal{T}} \rho(T)\ \underset{x}{\text{maximize}}\ \langle \theta(T), \phi(x) \rangle \quad \text{subject to } \mathbb{E}_\rho[\theta(T)] = \theta. \tag{2.50}$$

The tightest such bound is exactly given by the MAP LP relaxation (2.45) and is independent of the spanning tree distribution ρ. This surprising result can be shown by straightforward derivation of the Lagrangian dual and noting that strong duality holds [168]. Kolmogorov showed that this result holds more generaly for non-spanning trees so long as every edge is contained in at least one tree [93].

Reweighted Max-Product BP

Finding the tightest bound (2.50) is as difficult as solving the MAP LP relaxation (2.45). Reweighted max-product (RMP) tries to find a tight bound via message passing, which is often much more efficient than solving the dual LP directly [178, 179]. In the remainder of this section we make use of the following shorthand notation,

$$\theta_s(x_s) = \sum_{i=1}^{K} \theta_{s;i}\delta_i(x_s), \qquad \theta_{st}(x_s, x_t) = \sum_{i=1}^{K}\sum_{j=1}^{K} \theta_{st;ij}\delta_i(x_s)\delta_j(x_t) \tag{2.51}$$

With some algebra we can rewrite the RMP bound (2.49) in terms of *edge appearance* probabilities ρ_{st} for each edge $(s,t) \in \mathcal{E}$,

$$\underset{x}{\text{maximize}}\ \langle \theta, \phi(x) \rangle \leq \underset{x}{\text{maximize}} \left\{ \sum_{s \in \mathcal{V}} \theta_s(x_s) + \sum_{(s,t) \in \mathcal{E}} \rho_{st}\theta_{st}(x_s, x_t) \right\}. \tag{2.52}$$

Reweighted MP messages and pseudo-max-marginals are given by,

$$m_{ts}(x_s) = \underset{x_t}{\text{maximize}}\ \psi_t(x_t)\psi_{st}(x_s,x_t)^{\frac{1}{\rho_{st}}} \frac{\prod_{u\in\Gamma(t)\setminus s} m_{ut}(x_t)^{\rho_{ut}}}{m_{st}(x_t)^{1-\rho_{st}}} \tag{2.53}$$

$$\nu_s(x_s) \propto \psi_s(x_s) \prod_{u\in\Gamma(s)} m_{us}(x_s)^{\rho_{us}}, \tag{2.54}$$

where $\psi(x) = \exp(\theta(x))$ for the discrete MRF case we consider. Pseudo-max-marginals can also be computed for the pairwise terms $\nu_{st}(x_s, x_t)$ as a function of the messages. It remains to show conditions under which an RMP fixed point solves the MAP LP relaxation. Consider the case where pseudo-max-marginals contain a set of maximizers consistent across nodes and edges,

$$x_s^* \in \arg\max_{x_s}\ \nu_s(x_s), \qquad (x_s^*, x_t^*) \in \arg\max_{x_s,x_t}\ \nu_{st}(x_s,x_t). \tag{2.55}$$

Wainwright et al. [168] showed that this condition, known as *strong tree agreement* (STA), is sufficient for the RMP fixed point to solve the MAP LP relaxation. Moreover, if STA holds then the MAP LP relaxation (2.45) is tight and x^* in (2.55) is a MAP configuration. Weiss et al. [171] also showed that it is sometimes possible to construct a MAP configuration when STA does not hold, but Kolmogorov showed by counterexample that this correspondence does not always hold [93] for non-binary models. Kolmogorov and Wainwright provide stronger connections between RMP fixed points and the MAP LP relaxation for binary models [94].

Chapter 3

Particle Max-Product Belief Propagation

Graphical models allow us to capture complex global phenomena by specifying simpler local interactions. When making statistical inferences, however, model complexity often results in difficult calculations. Such problems arise, for example, in computer vision applications that involve estimating and tracking articulated objects from images [158, 82, 150, 187, 188] or in the estimation of appearance features parameterized by continuous quantities [157]; in signal processing where core problems involve tracking acoustic contacts [155, 12], estimating signals that arise from arbitrary stochastic processes [30] or spatial reasoning among distributed sensors [80]; and in computational biology where models of protein dynamics [148, 26] and structure [137, 35] require sampling and optimizing complex energy functions.

Traditional sampling-based approaches to marginal and *maximum a posteriori* (MAP) inference are lacking in several aspects: methods, such as importance sampling or stochastic local search, do not exploit structure encoded in the graphical model resulting in inference that scales poorly with the problem dimension [6, 77]; sequential Monte Carlo (SMC), or particle filters, exhibit classic particle degeneracies over moderate time scales [30]; and simulated annealing (SA) requires a cooling schedule that is prohibitively slow in practice [70].

In this chapter we develop a particle-based max-product algorithm for MAP inference in continuous MRFs. The approach stochastically samples realizations of each random variable, known as *particles*, to approximate continuous message functions. Our algorithm, *diverse particle max-product* (D-PMP), combines the flexibility of sampling-based approaches with the efficiency of message passing. At each stage of the algorithm D-PMP maintains a diverse set of posterior mode hypotheses, capturing multiple local optima and thereby avoiding classic degeneracies associated with

particle filters. The integer program (IP) underlying *diverse particle selection* encourages diversity in the maintained hypotheses, without requiring tuning of application-specific distances among hypotheses. Moreover, the IP formulation is a submodular maximization, allowing efficient greedy optimization with optimality guarantees that preserve pseudo-max-marginal approximations.

3.1 Particle-Based Message Approximations

A key benefit of message passing inference is that it efficiently enumerates realizations of each random variable to marginalize or maximize over joint configurations (c.f. Sec. 2.3). Markov chain Monte Carlo (MCMC) sampling [6, 110], by contrast, samples a single joint configuration at each pass of the algorithm, which can lead to long mixing times. Sequential Monte Carlo (SMC) maintains several joint configurations at each stage, but portions of trajectories cannot be interchanged as in dynamic programming. Furthermore, SMC trajectories are highly correlated, leading to well-known degeneracies over moderate time scales [30, 83, 69, 90]. Particle filters and MCMC can be combined to resolve some of these shortcomings, thereby leading to methods that iteratively improve SMC estimates [27, 125] or Metropolis-Hastings proposals based on particle filters [7]. Stochastic local search [77, 67] offers an alternative metaheuristic for MAP inference that avoids restrictions imposed on valid MCMC samplers, but it too iterates in the space of joint configurations.

In this section we review particle-based approximations for message passing inference. Continuous message functions are approximated via stochastic samples, or *particles*, which are resampled at each iteration. For marginal inference (Sec. 3.1.1) sum-product message integrals are approximated with importance sampling [79] or via Gaussian kernel density estimation [157, 82]. For MAP inference (Sec. 3.1.2) the continuous maximization in max-product messages is approximated via stochastic local search whereby particles are sampled and discarded at each iteration to discover modes of the distribution.

3.1.1 Sum-Product Particle BP

For simplicity we focus on a pairwise Markov random field (MRF) with graph $G = (\mathcal{V}, \mathcal{E})$, vertices $s \in \mathcal{V}$ and edges $(s,t) \in \mathcal{E}$, and density:

$$p(x) = \frac{1}{Z} \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t). \tag{3.1}$$

The sum-product variant of BP (c.f. Sec. 2.3.2) computes marginal densities via local message recursions. The local *belief* $q_s(x_s)$ and messages $m_{ts}(x_s)$ are given by:

$$q_s(x_s) \propto \psi_s(x_s) \prod_{t \in \Gamma(s)} m_{ts}(x_s), \quad m_{ts}(x_s) = \int \psi_t(x_t)\psi_{st}(x_s, x_t) \prod_{u \in \Gamma(t) \setminus s} m_{ut}(x_t)\, dx_t,$$

where $\Gamma(s) = \{t \mid (s,t) \in \mathcal{E}\}$ is the set of nodes neighboring s, and $\mathcal{X}_t$ is the continuous domain of x_t. The continuous BP updates do not directly provide a realizable algorithm as the integral over $\mathcal{X}_t$ may be intractable, and the message function $m_{ts}(x_s)$ may not have an analytic form.

Importance Sampling

BP message updates can be viewed as an expectation of the pairwise potential function $\psi_{st}(x_s, x_t)$. Importance sampling [6] provides an approximation of expectations via weighted samples:

$$\mathbb{E}[g(x)] = \int_{\mathcal{X}} g(x)p(x)\, dx \approx \sum_{i=1}^{N} g(x^{(i)})w(x^{(i)}), \tag{3.2}$$

$$x^{(i)} \sim q(x), \quad w(x) \propto \frac{p(x)}{q(x)}, \quad \sum_{i=1}^{N} w(x^{(i)}) = 1.$$

We draw N i.i.d samples $\{x^{(i)}\}_{i=1}^{N}$ from the *proposal distribution* $q(x)$. These samples represent weighted point masses in the *empirical measure* $\hat{p}(x) = \sum_{i=1}^{N} w(x^{(i)})\delta_{x^{(i)}}(x)$. The Monte Carlo estimate of (3.2) is then,

$$\hat{\mathbb{E}}[g(x)] = \sum_{i=1}^{N} g(x^{(i)})w(x^{(i)}). \tag{3.3}$$

Assuming $\hat{\mathbb{E}}[g(x)]$ exists and is finite, and that $q(x)$ is absolutely continuous w.r.t. $p(x)$, the estimator (3.3) is unbiased and consistent [64]. The best proposal distribution is one that minimizes the variance and is $q(x) = |g(x)|p(x)$. The name importance sampling comes from this result, namely that particles should be concentrated in areas where $p(x)$ and $|g(x)|$ are mutually large, and thus *important*.

For the case where $p(x)$ is known only up to a normalization constant importance sampling provides the following estimator,

$$\hat{\mathbb{E}}[g(x)] = \frac{\frac{1}{N}\sum_{i=1}^{N} g(x^{(i)})w(x^{(i)})}{\frac{1}{N}\sum_{i=1}^{N} w(x^{(i)})} = \sum_{i=1}^{N} g(x^{(i)})\widetilde{w}(x^{(i)}), \tag{3.4}$$

where $\widetilde{w}(x^{(i)})$ is a normalized importance weight. The estimator (3.4) is consistent, but biased since it is a ratio of two unbiased estimators [6].

Particle Belief Propagation

Particle BP uses importance sampling to approximate the continuous BP message updates [79]. Given particles $\mathbb{X}_t = \{x_t^{(i)}\}_{i=1}^N$ sampled from some proposal distribution $x_t^{(i)} \sim q_t$ the message approximation via importance sampling is:

$$\hat{m}_{ts}(x_s) = \sum_{i=1}^{N} \psi_{st}(x_s, x_t^{(i)}) w_t(x_t^{(i)}). \tag{3.5}$$

The importance weight for a sample $x_t^{(i)} \in \mathcal{X}_t$ compensates for the mismatch between the proposal and the true posterior,

$$w_t(x_t^{(i)}) = \frac{\psi_t(x_t^{(i)}) \prod_{u \in \Gamma(t) \setminus s} \hat{m}_{ut}(x_t^{(i)})}{q_t(x_t^{(i)})}. \tag{3.6}$$

We can approximate the continuous BP beliefs $\hat{q}_s(x_s)$ over the particles $\mathbb{X}$ by substituting the message approximations $\hat{m}$ from Eq. (3.5). It can be beneficial to sample particles from the approximate marginals using a Metropolis-Hastings MCMC sampler to iteratively draw proposals [92, 79].

For junction tree representations of Bayesian networks Koller et al. [92] describe a general framework for approximating clique marginals. The nonparametric BP [157] and PAMPAS [82] algorithms approximate continuous BP messages with kernel density estimates, and use Gibbs samplers [81] to propose particles from belief distributions. The sum-product *particle belief propagation* (PBP) algorithm of [79] associates particles with nodes rather than messages or cliques, and thus avoids the need for explicit marginal density estimates.

3.1.2 Particle Max-Product

The max-product (MP) variant of BP is similar to the sum-product form, where messages maximize, instead of marginalize, over joint configurations (see Sec. 2.3.3). For tree-structured MRFs the beliefs $q_s(x_s)$ correspond to *max-marginal* distributions [168], which encode the probability of the most likely joint configuration,

$$q_s(x_s) \propto \underset{x'}{\text{maximize}}\ p(x') \ \text{ subject to } x'_s = x_s. \tag{3.7}$$

For MRFs with cycles the reweighted max-product (RMP) algorithm approximates max-marginal distributions via message passing (see Section 2.3.4). Given a spanning tree distribution with *edge appearance probabilities* ρ_{st} the RMP message

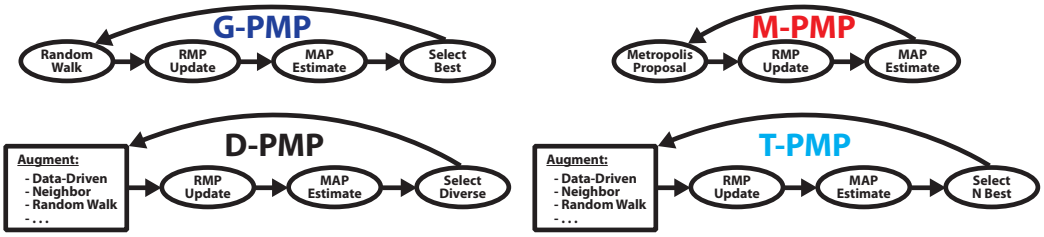


Figure 3.1: **Loopy PMP flowcharts.** Each variant of PMP has several stages: the particle set is augmented with draws from stochastic proposals, then RMP messages are updated on the augmented particles, and finally a subset of particles are discarded to control computation. Different particle selection methodologies lead to different PMP methods, such as Greedy PMP (G-PMP) of [128], the Top-N PMP (T-PMP) of [23], the Metropolis PMP (M-PMP) of [95], and the Diverse PMP (D-PMP) of [123].

update and pseudo-max-marginal are given by,

$$\widetilde{m}_{ts}(x_s) = \underset{x_t \in \mathcal{X}_t}{\text{maximize}}\ \psi_t(x_t)\psi_{st}(x_s, x_t)^{\frac{1}{\rho_{st}}} \frac{\prod_{u \in N(t)\setminus s} \widetilde{m}_{ut}(x_t)^{\rho_{ut}}}{\widetilde{m}_{st}(x_t)^{1-\rho_{st}}} \tag{3.8}$$

$$\widetilde{\nu}_s(x_s) \propto \psi_s(x_s) \prod_{u \in \Gamma(s)} \widetilde{m}_{us}(x_s)^{\rho_{us}} \approx q_s(x_s). \tag{3.9}$$

The messages $\widetilde{m}$ involve a continuous maximization, which may have no closed-form solution and no compact representation. Particle max-product (PMP) methods approximate these continuous functions by optimizing over a discrete set of *particles* $\mathbb{X} \subset \mathcal{X}$ found via stochastic search. Each iteration monotonically increases a lower bound on the true MAP probability:

$$\underset{x \in \mathbb{X}}{\text{maximize}}\ \log p(x) \leq \underset{x \in \mathcal{X}}{\text{maximize}}\ \log p(x). \tag{3.10}$$

Each PMP iteration improves this bound in several stages, summarized in Fig. 3.1, which we describe in detail in the following sections.

Augment via Stochastic Proposals

Given a current set of N particles $\mathbb{X}_t \subset \mathcal{X}_t$ a stochastic local search seeks higher-likelihood configurations. At each iteration PMP first creates an augmented particle set $\mathbb{X}^{\text{aug}} = \mathbb{X} \cup \mathbb{X}^{\text{prop}}$ of size αN for $\alpha > 1$. New particles are drawn from proposal distributions $\mathbb{X}^{\text{prop}} \sim q(\mathbb{X})$. In the simplest case, Gaussian random walk proposals $q^{\text{gauss}}(x_s) = N(x_s \mid \bar{x}_s, \Sigma)$ sample perturbations of current particle locations $\bar{x}_s$ [163, 128]. For some models, a more informative *neighbor-based* proposal is possible that samples from edge potentials $q^{\text{nbr}}(x_s \mid \bar{x}_t) \propto \psi_{st}(x_s, \bar{x}_t)$ conditioned on a particle $\bar{x}_t$ at neighboring node $t \in \Gamma(s)$ [23]. Specialized "bottom-up" or "data-driven" proposals based on approximations of observation potentials $\psi_s(x_s)$ can also be effective [123].

Reweighted Max-Product Optimization

Standard or reweighted MP message updates are used to approximate the max-marginal distribution of each proposed particle. The αN values of each discrete message vector satisfy

$$m_{ts}(x_s) = \max_{x_t \in \mathbb{X}_t^{\text{aug}}} \psi_t(x_t)\psi_{st}(x_s, x_t)^{\frac{1}{\rho_{st}}} \frac{\prod_{u \in \Gamma(t) \setminus s} m_{ut}(x_t)^{\rho_{ut}}}{m_{st}(x_t)^{1-\rho_{st}}}. \tag{3.11}$$

Note that messages are computed by discrete optimization over $\mathbb{X}^{\text{aug}}$, the augmented particle set. Pseudo-max-marginals can be computed via the approximate messages, $\nu_s(x_s) \propto \psi_s(x_s) \prod_{u \in \Gamma(s)} m_{us}(x_s)^{\rho_{us}}$. Message updates require $\mathcal{O}(\alpha^2 N^2)$ operations, and compute the pseudo-max-marginal $\nu_s(x_s)$ for each $x_s \in \mathbb{X}^{\text{aug}}$.

Particle Selection

Particles are accepted or rejected to yield N new states $\mathbb{X}^{\text{new}} \subset \mathbb{X}^{\text{aug}}$. Particle selection makes subsequent iterations more computationally efficient by avoiding an unbounded growth in the number of particles, but can lead to classic degeneracies if done improperly. Several different particle selection methods have been proposed, which we briefly review.

The simple *greedy PMP* (G-PMP) method selects the single particle with the highest max-marginal value $x_s^* = \arg\max_{x_s \in \mathbb{X}_s^{\text{aug}}} \nu_s(x_s)$, and samples particles from Gaussian random walk proposals $q^{\text{gauss}}(x_s)$ with mean x_s^* [163, 128]. G-PMP updates are computationally efficient, but the greedy selection does not retain particles near multiple modes and the random walk proposals do not effectively explore high-dimensional spaces.

A less greedy selection method is *top-N PMP* (T-PMP), which retains the N particles with the highest estimated max-marginal probability. This *top-N PMP* [123] generalizes PatchMatch BP [23], a method specialized to low-level vision tasks which utilizes top-N particle selection and neighbor proposals. T-PMP finds high probability solutions quickly, but the top-N particles are often slight perturbations of the same solution, reducing the number of effective particles and causing sensitivity to initialization.

Building directly on the sum-product PBP algorithm of [79], Kothapa et al. [95] proposed *Metropolis PMP* (M-PMP), which approximately samples particles from the current max-marginal estimate using a Metropolis sampler with Gaussian random walk proposals. Because the entire particle set is replaced at each iteration, discovered modes may be lost and the bound of Eq. (3.10) may decrease. While drawing particles from max-marginals does explore important parts of the state space, computing the

Metropolis acceptance-ratio requires an expensive $\mathcal{O}(N^2)$ message update. Moreover, because we do not seek to approximate expectations as in PBP, the bias concerns that motivate traditional importance sampling methods do not apply here.

3.2 Diverse Particle Max-Product

To avoid degeneracies common to other methods the D-PMP

particles via a *diverse selection* step favoring states which minimally distort the current RMP messages. The algorithm, outlined in Fig. 3.1, naturally encourages diversity by preserving solutions near multiple local optima thereby enabling D-PMP to reason more globally than other methods.

Diverse PMP discards states that are not necessary to retain accurate message approximations by selecting the subset of particles that minimize distortion from the RMP messages m. In Section 3.2.1 we present a distortion measure that yields a submodular optimization, with an efficient greedy approximation and a multiplicative optimality bound. We also consider a formulation based on minimizing maximum message distortions (Sec. 3.2.2) which has shown good empirical results, but analysis is more difficult than for the submodular formulation.

3.2.1 Diverse Particle Selection

For each node $t \in \mathcal{V}$ we select a subset of particles via the indicator vector $z \in \{0,1\}^{\alpha N}$, where $z(i) = 1$ denotes that particle $x^{(i)} \in \mathbb{X}^{\text{aug}}$ is selected. The message vector over this subset is $\hat{m}_{ts}(z)$ and messages m_{ts} over all particles $\mathbb{X}^{\text{aug}}$ are given by,

$$m_{ts}(a) = \underset{b\in\{1,\dots,\alpha N\}}{\text{maximize}}\ M_{st}(a,b), \qquad \hat{m}_{ts}(a;z) = \underset{b\in\{1,\dots,\alpha N\}}{\text{maximize}}\ z(b)M_{st}(a,b). \tag{3.12}$$

For notational convenience we have combined terms needed for RMP message updates into a *message foundation* matrix $M_{st} \in \mathbb{R}^{\alpha N\times\alpha N}$,

$$M_{st}(a,b) = \psi_t(x_t^{(b)})\psi_{st}(x_s^{(a)}, x_t^{(b)})^{\frac{1}{\rho_{st}}} \frac{\prod_{u\in\Gamma(t)\setminus s} m_{ut}(b)^{\rho_{ut}}}{m_{st}(b)^{1-\rho_{st}}}. \tag{3.13}$$

We choose the subset of particles that minimizes *total* distortion between the messages $\hat{m}_{ts}(z)$ and m_{ts}, resulting in the following integer program (IP):

$$\begin{aligned} &\underset{z}{\text{minimize}} \sum_{s\in\Gamma(t)} \sum_{a=1}^{\alpha N} \left(m_{ts}(a)^{\rho_{ts}} - \hat{m}_{ts}(a;z)^{\rho_{ts}} \right) \\ &\text{subject to} \quad \|z\|_1 \le N, \quad z \in \{0,1\}^{\alpha N}. \end{aligned} \tag{3.14}$$

By incorporating edge appearance probabilities ρ_{st}, which scale the relative height of modes, D-PMP more accurately preserves messages over edges that are well-represented in the set of spanning trees.

Submodularity and Other Properties of Diverse Selection

The objective (3.14) encourages states which preserve RMP pseudo-max-marginal approximations. For each node $s \in \mathcal{V}$ the pseudo-max-marginal ν_s is a product of incoming messages from neighbors $t \in \Gamma(s)$, allowing us to link message distortion to pseudo-max-marginal distortion:

Proposition 3.2.1 *Let* $0 \preceq \hat{m} \preceq m \preceq 1$, $0 \leq \psi \leq 1$ *and edge appearance probabilities* $\rho_{st} \in [0, 1]$. *For all nodes* $s \in \mathcal{V}$ *we have:*

$$\|\nu_s - \hat{\nu}_s\|_1 \leq \sum_{t \in \Gamma(s)} \sum_{x_s} \Big[m_{ts}(x_s)^{\rho_{ts}} - \hat{m}_{ts}(x_s)^{\rho_{ts}} \Big]. \tag{3.15}$$

The intuition is the following: discarding particles introduces error into the messages and pseudo-max-marginals, and the diverse selection IP (3.14) minimizes the upper bound (3.15) ensuring that pseudo-max-marginal approximations are preserved. In addition to bounding the pseudo-max-marginal error, we show that diverse selection corresponds to a submodular maximization.

Submodularity A set function $f : 2^Z \to \mathbb{R}$ defined over subsets of Z is submodular iff for any subsets $Y \subseteq X \subseteq Z$ and an element $e \notin X$ the function f satisfies,

$$f(Y \cup \{e\}) - f(Y) \geq f(X \cup \{e\}) - f(X).$$

The property of submodularity states that adding an element e to the smaller set Y produces more gain than adding it to the larger set X. This property of *diminishing marginal gain* is formalized by the quantity $\Delta(Y, e) \triangleq f(Y \bigcup \{e\}) - f(Y)$, known as the *margin*.

Proposition 3.2.2 *The optimization* (3.14) *is equivalent to maximizing a monotonic submodular objective subject to cardinality constraints.*

The result of Prop. (3.2.2) follows by reformulating the objective (3.14) as a facility location problem, which is submodular [120]. In Appendix A.2 we provide a constructive proof that does not rely on earlier results. The selection IP (3.14) is NP-hard [36], and in the next section we develop a well-known greedy approximation for this problem which obtains a $(1 - 1/e)$ multiplicative optimality bound [115, 106, 120].

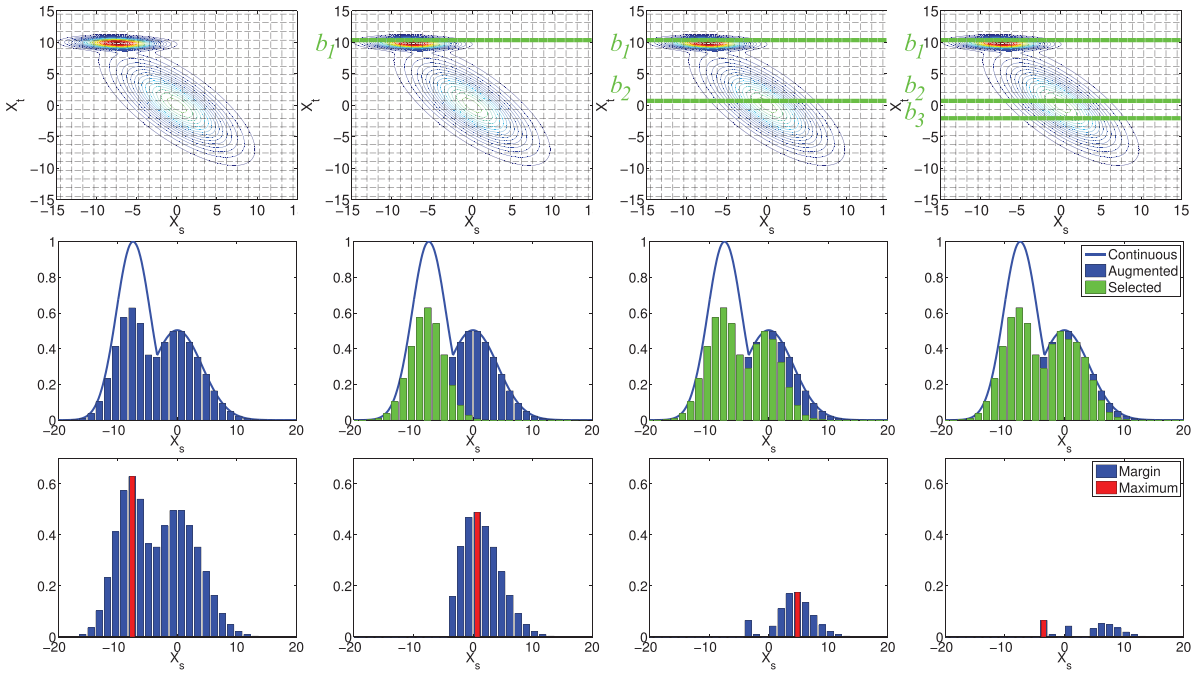


Figure 3.2: **Lazy greedy particle selection** of x_t to for a two-node correlated Gaussian model with compatibility $\psi_{st}(x_s, x_t) = N(x \mid \mu_{st}, \Sigma_{st})$ and unary potentials of evenly-weighted mixtures of two Gaussians. Particles are greedily selected left-to-right. *Top Row:* Message foundation matrix showing particle set as a regular grid (dashed lines) to aid visualization and selected particles (green). *Middle Row:* Message over augmented particles m_{ts} (blue) and subset $\hat{m}_{ts}$ (green). We also show the continuous message $\widetilde{m}_{ts}$ for reference (blue line). *Bottom Row:* Margin at each particle selection (blue) with selected particle margin (red).

Lazy Greedy Selection

The standard *lazy greedy* approach to submodular maximization exploits diminishing marginal returns to avoid redundant computations [115, 106]. Each iteration updates and sorts the largest margin until a stable maximizer is found. The algorithm terminates when N particles are selected, or the maximum margin is zero. In this section we formulate the lazy greedy algorithm for particle selection.

Initialize: For each node t let $M = \left[(M_{s_1 t}^{\rho_{s_1 t}})^T, \ldots, (M_{s_d t}^{\rho_{s_d t}})^T\right]^T$ be the reweighted message foundations of neighbors $\Gamma(t) = \{s_1, \ldots, s_d\}$ as in Eq. (3.13). Initialize the selection vector z and margins:

$$\Delta(b) = \sum_{a=1}^{d\alpha N} M(a, b), \quad z(b) = 0, \quad \forall\, b \in \{1, \alpha N\}. \tag{3.16}$$

First Iteration: Ensure that the current MAP estimate x^* is never discarded by

setting $z(b^*) = 1$, where b^* is the index of x_t^* in the augmented particle set $\mathbb{X}_t^{\text{aug}}$. Update the message approximation $\hat{m}(a) = M(a, b^*)$.

Iterations 2 to N: Choose the largest margin to update $\widetilde{b} = \arg\max_{\{b|z(b)=0\}} \Delta(b)$. If $\Delta(\widetilde{b}) = 0$ then terminate prematurely since the message can be perfectly reconstructed with a subset of particles. If $\Delta(\widetilde{b})$ has already been updated on the current iteration then set $z(\widetilde{b}) = 1$ and update the message approximation $\hat{m}(a) = \max(\hat{m}(a), M_t(a, \widetilde{b}))$. Otherwise, update the margin and repeat,

$$\Delta(\widetilde{b}) \triangleq \sum_a \Big[\max(\hat{m}(a), M(a, \widetilde{b})) - \hat{m}(a)\Big]. \tag{3.17}$$

The lazy greedy algorithm iteratively updates and re-sorts margins from the previous iteration in decreasing order. A stable maximizer is achieved when a margin remains the maximizer after the update (3.17). Margins are guaranteed to be non-increasing as new particles are considered, and so this partial ordering is guaranteed to find the selection that produces the largest marginal gain. While worst-case computation is $\mathcal{O}(\alpha N^2)$, in practice only a few margins are typically updated to find the maximizer, thus avoiding quadratic complexity. Moreover, computation can be dramatically reduced by storing margins in a max heap, allowing $\mathcal{O}(1)$ access to the maximizer. Figure 3.2 graphically demonstrates lazy greedy particle selection for a simple two-node Gaussian mixture model.

3.2.2 Minimax Particle Selection

Another possible distortion measure is to minimize the maximum message error. This approach prefers particles to approximate local maxima via the following IP:

$$\begin{aligned} &\underset{z}{\text{minimize}} \ \underset{s\in\Gamma(t),\, 1\le a\le \alpha N}{\text{maximize}} \ m_{ts}(a)^{\rho_{st}} - \hat{m}_{ts}(a; z)^{\rho_{st}} \\ &\text{subject to} \quad \|z\|_1 \le N, \quad z \in \{0,1\}^{\alpha N}. \end{aligned} \tag{3.18}$$

Like the submodular formulation this IP is NP-hard and so we present a greedy approximation algorithm.

Following a development similar to the lazy greedy algorithm (Sec. 3.2.1) for each node t let $M = \left[(M_{s_1 t}^{\rho_{s_1 t}})^T, \ldots, (M_{s_d t}^{\rho_{s_d t}})^T\right]^T$ be the message foundations of neighbors $\Gamma(t) = \{s_1, \ldots, s_d\}$ as in Eq. (3.13). Maximizing over the columns of $M \in \mathbb{R}^{d\alpha N \times \alpha N}$ produces a concatenated vector of outgoing messages to all neighbors. Similarly, maximizing over any subset of columns (indexed by $z \in \{0,1\}^{\alpha N}$) produces a concatenated vector of messages computed on a subset of the corresponding particles,

$$m(a) = \underset{1\le b\le \alpha N}{\text{maximize}} \ M(a, b), \quad \hat{m}(a; z) = \underset{1\le b\le \alpha N}{\text{maximize}} \ z(b) M(a, b), \quad \forall 1 \le a \le d\alpha N.$$

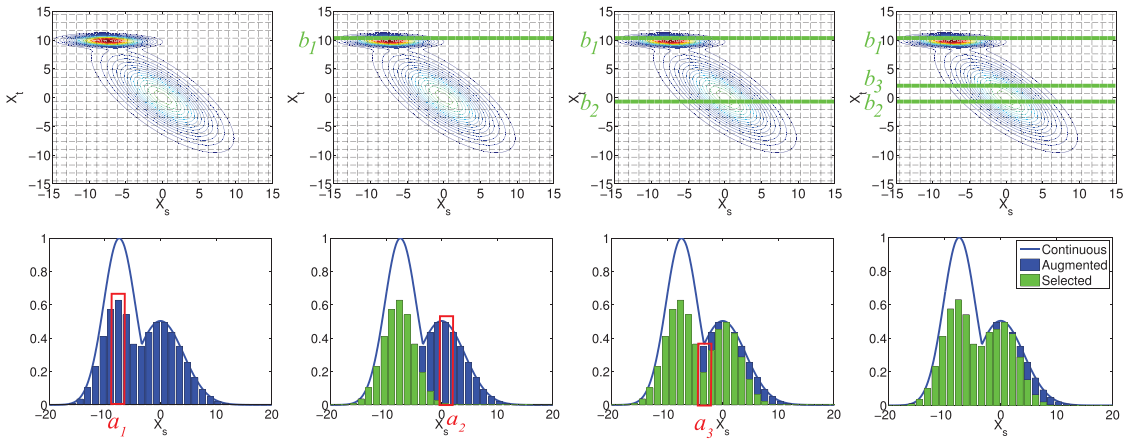


Figure 3.3: **Minimax particle selection.** Three particles selected from left-to-right on node x_t for the Gaussian mixture model in Fig. 3.2. *Top Row:* Level sets of the foundation matrix showing regular grid of particles (dashed black) and selected particles b_k (green). *Bottom Row:* Message m_{ts} over augmented particles (blue) with selected $\arg\max$ location a_k (red) and $\hat{m}_{ts}$ over subset (green). Continuous message (blue line) is computed by numerical approximation for reference.

Starting with an empty particle set $z^{(0)} = \vec{0}$ at iteration $k = 1, \ldots, N$ select a single particle with index b_k, and update the selection $z^{(k)}(b_k) = 1$. Each step improves our approximation of the augmented messages m, and because maximization is associative we can incrementally update the approximation,

$$\hat{m}(a; z^{(k)}) = \max\left\{\hat{m}(a; z^{(k-1)}),\ M(a, b_k)\right\}. \tag{3.19}$$

To choose the next particle identify the index $a_k \in \{1, \ldots, d\alpha N\}$ with the largest distortion and select the particle index $b_k \in \{1, \ldots, \alpha N\}$ that minimizes this error:

$$a_k = \underset{1 \le a \le d\alpha N}{\arg\max}\ \ m(a) - \hat{m}(a; z^{(k-1)}), \quad b_k = \underset{1 \le b \le \alpha N}{\arg\max}\ \ M(a_k, b), \tag{3.20}$$

and set $z^{(k)}(b_k) = 1$. The particle selection of Eq. (3.20) always eliminates errors in the max-product message for particle a_k, and may also reduce or eliminate errors in messages for particles a where $\psi_{st}(x_s^{(a)}, x_t^{(b_k)})$ is large.

Each step of the greedy algorithm requires $\mathcal{O}(d\alpha N)$ time, so the overall cost of selecting N particles is $\mathcal{O}(d\alpha N^2)$. This quadratic cost is comparable to the RMP message updates in Eq. (3.11). While our experiments treat N as a fixed parameter trading off accuracy with computational cost, it may be useful to vary the number of selected particles across nodes or iterations of D-PMP, for example by selecting particles until some target error level is reached. See Figure 3.3 for a graphical depiction of the greedy selection procedure on the toy Gaussian mixture model.

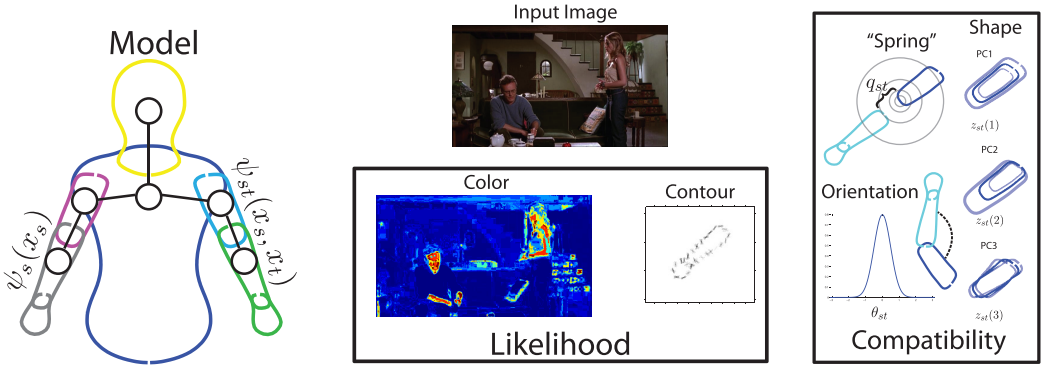


Figure 3.4: **Human pose estimation** *Deformable Structures* [187] models human pose and shape with a loose-limbed Gaussian prior. *Spring* potentials encourage connected body components, while a PCA model encodes part shape. A contour-based likelihood scores part location in the image via an SVM classifier, trained on HOG features for each body part, and a logistic regression function calibrates scores across each part [129]. A skin color likelihood measures appearance similarity of the lower arms to a histogram of skin color.

3.3 Experimental Results

Particle max-product imposes few restrictions on the MRF structure, allowing it to be applied in many applications. The approach is well suited to models of articulated physical objects, such as when reasoning about human figures in images. In the following sections we consider human pose estimation in both images (Sec. 3.3.1) and video (Sec. 3.3.2). We conclude with an application to optical flow estimation for images pairs (Sec. 3.3.3).

3.3.1 Single Image Human Pose Estimation

We model human pose and shape in single images using the *deformable structures* (DS) model [187], an articulated part-based human body representation. Unlike the related Pictorial Structures (PS) model [50], high-dimensionality of the DS state space makes discretization infeasible.

Deformable Structures

The DS model specifies a pairwise MRF with nodes $s \in \mathcal{V}$ for each body part, and links kinematic neighbors with edges $(s,t) \in \mathcal{E}$ (Figure 3.4). With global rotation θ_s, scale d_s, center o_s, and shape z_s, the state of part s is

$$x_s = (z_s, o_s, \sin(\theta_s), \cos(\theta_s), d_s)^T. \tag{3.21}$$

Shape is modeled via PCA analysis of part-specific training data; we learn a transformation matrix B_s and mean m_s for each part type. A set of contour points c_s and

joint locations p_s are given in object-centered coordinates via a linear mapping,

$$(c_s, p_s)^T = B_s z_s + m_s. \tag{3.22}$$

The likelihood of pose x_s is obtained by projecting these joint locations into image coordinates with the rotation matrix $R(\theta_s)$, scaling d_s, and translation $t(o_s)$,

$$i_s(x_s) = d_s R(\theta_s) \begin{pmatrix} c_s \\ p_s \end{pmatrix} + t(o_s). \tag{3.23}$$

Using the above representation the DS joint probability is a pairwise MRF involving three types of potentials:

$$p(x) \propto \prod_{s \in \mathcal{V}} \psi_s^{\text{contour}}(x_s) \psi_s^{\text{skin}}(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}^{\text{body}}(x_s, x_t). \tag{3.24}$$

Image likelihoods are given by two complementary potentials which capture information about boundary contours and skin color. The contour likelihood is based on an SVM classifier trained on *histogram of oriented gradients* (HOG) features $h_s(i_s(x_s))$ [40]. SVM scores $f_s(h_s(i_s(x_s)))$ are mapped to calibrated probabilities via logistic regression [129], using a weight a_s and bias b_s learned from validation data:

$$\psi_s^{\text{contour}}(x_s) = \frac{1}{1 + \exp(a_s f_s(h_s(i_s(x_s))) + b_s)}. \tag{3.25}$$

The skin color likelihood $\psi_s^{\text{skin}}(x_s)$ captures the tendency of lower arms to be unclothed, and is derived from a histogram model of skin appearance [187].

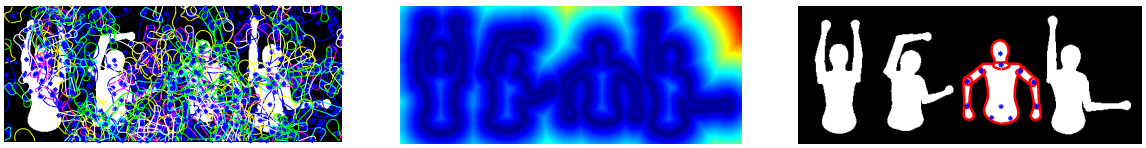
The kinetic prior between a pair of neighboring body parts captures relative displacement, orientation and scale difference. Neighboring parts $(s,t) \in \mathcal{E}$ are connected by joints with locations p_{st} and p_{ts}, respectively. The relative displacement $q_{ts} = p_{ts} - p_{st}$, relative orientation $\theta_{ts} = \theta_t - \theta_s$, and scale difference $d_{ts} = d_t - d_s$ are computed via the transformation,

$$T_{st}(x_s, x_t) = (z_s, z_t, \sin(\theta_{ts}), \cos(\theta_{ts}), q_{ts}, d_{ts})^T. \tag{3.26}$$

The prior distribution models relative position, orientation, and scale via a multivariate Gaussian,

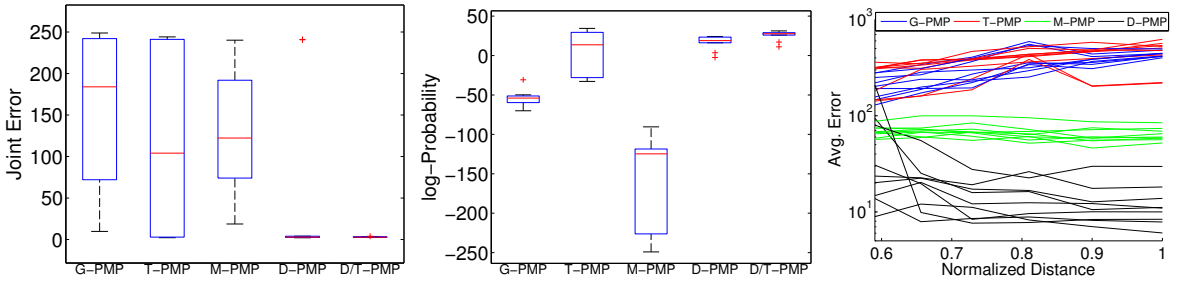
$$\psi_{st}^{\text{body}}(x_s, x_t) \propto N(T_{st}(x_s, x_t) \mid \mu_{st}, \Sigma_{st}) \mathbb{I}_{\mathcal{A}}(d_s, \theta_s) \mathbb{I}_{\mathcal{A}}(d_t, \theta_t), \tag{3.27}$$

where the indicator function $\mathbb{I}_{\mathcal{A}}(\cdot)$ enforces validity of angular components and nonnegativity of the scale parameters by the constraint set $\mathcal{A} = \{d, \theta \mid d > 0,\ \sin^2\theta + \cos^2\theta = 1\}$.



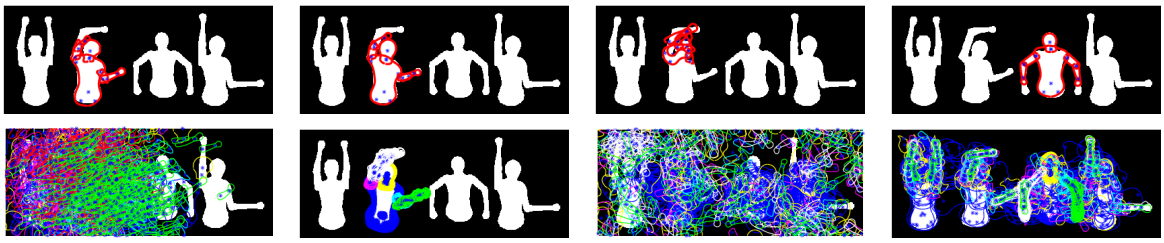
Initialization Contour Likelihood MAP

Figure 3.5: **Synthetic pose estimation.** *Left:* Initial particles sampled uniformly at random in the image plane. *Center:* Synthetic likelihood proportional to squared distance from image contours. *Right:* Ground truth MAP estimate corresponds to second-from-right figure.



Joint Error Log-Probability Joint Error vs. Dist.

Figure 3.6: **Synthetic image experiments** *Left:* Box plots for 10 trials of the "ICML" experiment, where the joint error equals the L_2 distance from the true MAP pose, averaged over all joints. *Center:* Log-probability of the most likely configuration identified by each method. *Right:* Average joint error in the *distance experiment* plotted versus the distance separating the 9 poses. Each line shows estimation error for a single pose across 6 images.



G-PMP T-PMP M-PMP D-PMP

Figure 3.7: **Typical synthetic pose estimation results** We show the final MAP estimate (top) and 200 particles per part (bottom) for each method.

Synthetic Images

We compare D-PMP with baseline methods on a set of synthetic images by using a simplified version of the DS model. The reduced model does not include skin likelihood potentials ψ^{skin} and the contour likelihood ψ^{contour} is based on squared distance from image edges, see Fig. 3.5. We conduct two experiments that evaluate accuracy of the MAP estimate and sensitivity to initialization for each PMP method.

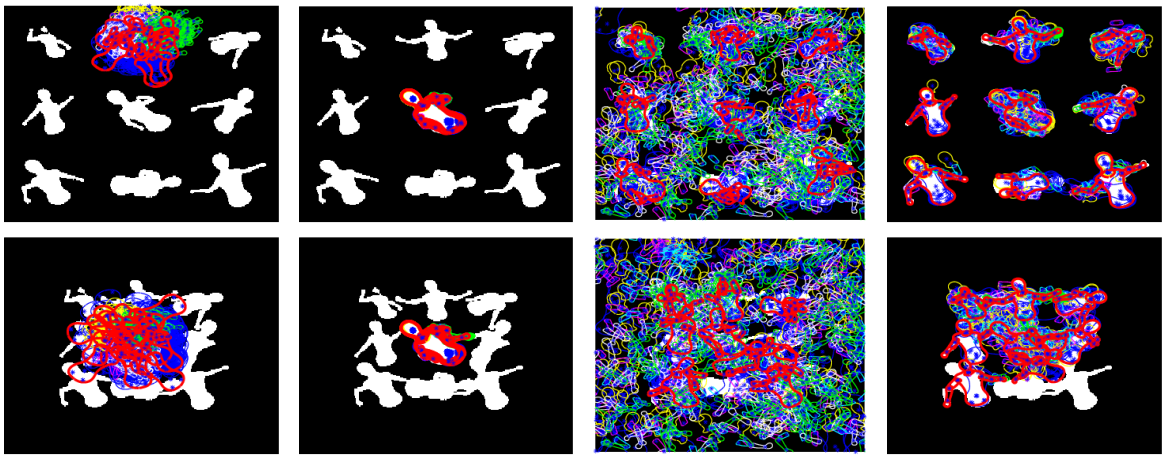


Figure 3.8: **Preserving multiple modes** Figures do not overlap at the furthest spacing (top), but extremities overlap at the closest spacing (bottom). Each method is run for 300 iterations from 30 random 200-particle initializations. The top 9 modes (red) are obtained by selecting the closest torso particle to each ground truth puppet, and from this a Viterbi backward pass generates the remaining limbs.

In all cases we run 300 PMP iterations with a total budget of 200 particles.

In the first experiment we use a hand-constructed image containing four silhouettes arranged to spell "ICML" (Figure 3.4). Using exhaustive gradient optimization we verify that the third figure from the left (the letter "M") corresponds to the global MAP. We run for 10 uniformly random initializations and report error of the predicted joint locations in Figure 3.6. D-PMP consistently produces MAP estimates near the global peak, and thereby produces lower error than other methods which are sensitive to initialization. We also consider a hybrid method, D/T-PMP, in which D-PMP is run for the first 200 iterations and T-PMP for the final 100 to refine the estimate and provide better alignment.

To evaluate diversity we sample 9 *puppets* from the DS model prior and arrange them in a 3×3 grid. We measure the ability of each PMP method to preserve hypotheses over a sequence of 6 images with decreasing relative distance between puppets (Figure 3.8). Using an oracle to select the torso particle closest to each puppet we generate a conditional mode of the remaining parts via a Viterbi-style backward pass. Figure 3.6 (right) plots average error versus puppet distance for each of the 9 puppets. D-PMP maintains significantly better mode estimates compared to other methods as shown by the final particles in Figure 3.8. We observe sensitivity to local optima in T-PMP and G-PMP, which generally capture only a single mode. M-PMP scatters particles widely, but does a poor job of concentrating particles on modes of interest.

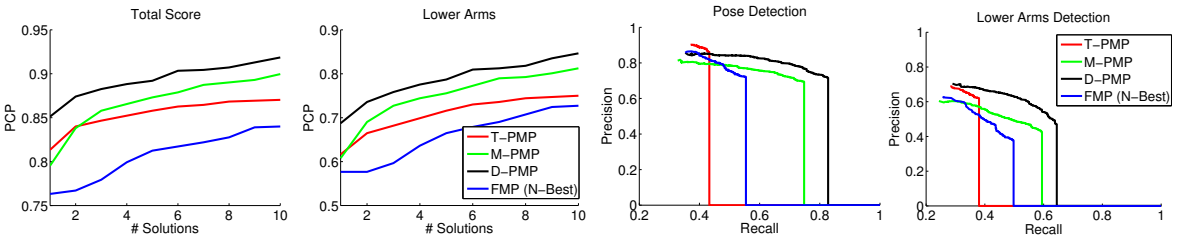
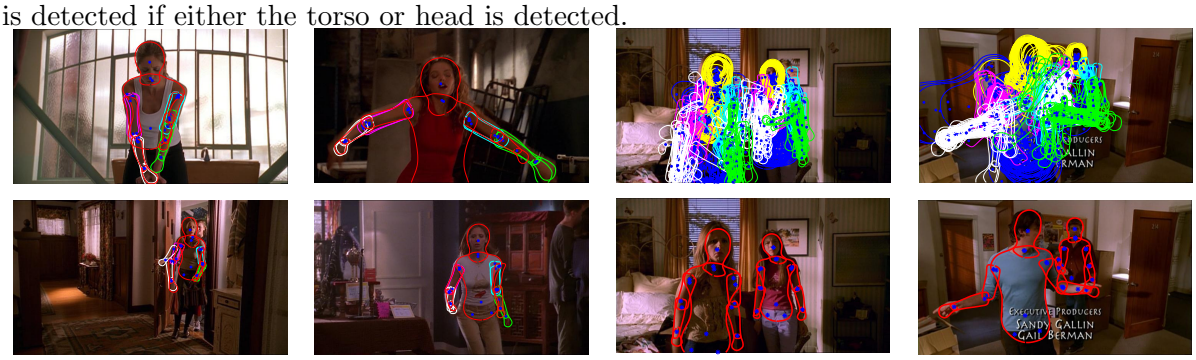


Figure 3.9: **Detection results.** *Left:* Average PCP score versus the number of hypotheses for images containing a single person. We report PCP for all parts and for only the lower arms. *Right:* Precision-recall curves for images containing multiple people. We report full body detections and lower arm detections, determined via a PCP threshold of 0.5. A body is detected if either the torso or head is detected.



Single Person **Multiple People**

Figure 3.10: **Preserving multiple hypotheses** *Left:* Single person images showing a MAP estimate (red) with poor arm placement. The second and third ranked solutions preserved by D-PMP, by max-marginal values, are shown for upper (magenta-cyan) and lower arms (white-green); they offer much greater accuracy. *Right:* The full set of particles at the final iteration of D-PMP shows multiple hypotheses retained about multiple people (top). For each person, we also plot the best pose in the set of retained hypotheses (bottom, red).

Real Images

Next, we look at real images from the *Buffy the Vampire Slayer* dataset [52] pose estimation benchmark. The dataset consists of a standard partition of 276 test images and nearly 500 training images. We use a recent set of *stickmen* annotations for all figures in the dataset [99] and report separate results on frames containing single and multiple figures.

Each inference method initialized using 100 particles sampled around candidate hypotheses from the *flexible mixture of parts* (FMP) pose estimation method [177], pruned below a score of 0.5 and followed by non-maximal suppression with overlap threshold 0.8. We run each method with 100 particles for 100 iterations and compare to the N-best maximal decoders computed on the raw FMP detections [126].

For single-person images we use the standard *percentage of correctly estimated parts* (PCP) distance-based detection metric for evaluation. Unlike results reported by Ferrari et al. [52] we normalize PCP by the fixed number of images in the dataset, thereby avoiding irregularities when varying the number of hypotheses. Pose hypotheses are sorted according to their max-marginal value (or FMP score), and Figure 3.9 shows detection accuracy versus the number of hypothesized poses. While accuracy for the arms are uniformly lower than total detection, the trends are is similar: given an identical model D-PMP is more accurate than conventional particle max-product algorithms. We offer qualitative examples of how D-PMP preserves alternative (upper and lower arm) hypotheses in Figure 3.10.

Figure 3.9 reports precision-recall (P-R) for multiple people, this the a standard metric for multiclass object detection. A body is considered detected if the torso or head PCP score is 1, and we evaluate the lower arm detection separately. Our P-R methodology differs slightly from those used in PSCAL VOC in that we compute detection considering the top scoring poses within each image, rather than a ranking across images. D-PMP again outperforms all other methods, both for body detection as well as for lower arm detection. Figure 3.10 offers qualitative examples of D-PMP's ability to preserve hypotheses about multiple people in an image. Without an explicit model of multiple people, we are able to infer their existence by finding multiple diverse posterior modes.

3.3.2 Articulated Pose Tracking in Video

Integrating information over a video sequence enables pose estimation that is robust to transient artifacts arising from motion blur, poor lighting, and other appearance variations. In practice, however, little benefit has been reported from modeling temporal dynamics, with most work showing superior accuracy by ignoring temporal dependence [52] or by modeling limited temporal dependence [139, 188]. These failures are largely the result of inadequate inference that is unable to exploit the rich structure imposed by a dynamical model of pose evolution. To address this we adopt the *flowing puppets* model of Zuffi et al. [188] and develop PMP inference to jointly infer pose over the video sequence (see Figure 3.11).

Flowing Puppets

The flowing puppets model builds on Deformable Structures by placing a prior over motion between consecutive frames. The likelihood model incorporates temporal information via optical flow estimates. Let the set of vertices $s \in \mathcal{V}$ is $|\mathcal{V}| = N \times T$

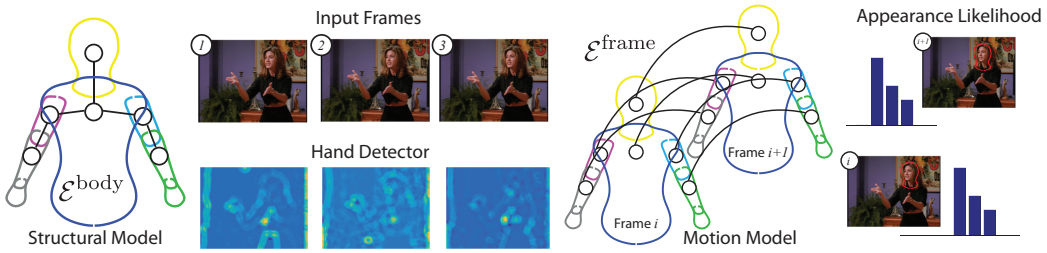


Figure 3.11: **Tracking human pose and shape.** The flowing puppets [188] model (left) extends the DS model of structural kinematics by incorporating a flow-based hand detector (center-left). Appearance constancy terms (right) enforce the notion that parts are largely invariant to significant changes in appearance over short time scales. We further extend the model with a Gaussian scale mixture motion model.

for N parts and T frames, the joint probability is then:

$$p(x) \propto \prod_{s \in \mathcal{V}} \psi_s^{\text{contour}}(x_s)\psi_s^{\text{skin}}(x_s)\psi_s^{\text{hand}}(x_s) \prod_{(s,t)\in\mathcal{E}^{\text{part}}} \psi_{st}^{\text{body}}(x_s, x_t) \prod_{(s,t)\in\mathcal{E}^{\text{frame}}} \psi_{st}^{\text{appearance}}(x_s, x_t)\psi_{st}^{\text{motion}}(x_s, x_t). \tag{3.28}$$

The MRF (3.28) is defined over sets of edges, within-frame edges $\mathcal{E}^{\text{part}}$ and edges between identical parts in consecutive frames $\mathcal{E}^{\text{frame}}$. A graphical depiction of edge sets is shown in Figure 3.11.

The the structural prior ψ_{st}^{body} is identical the DS prior (3.24) as are the skin color ψ_s^{skin} and contour likelihoods ψ_s^{contour}. A motion model encodes the relative displacement of part joint locations in neighboring frames. For a pair of nodes $(s,t) \in \mathcal{E}^{\text{frame}}$ the relative displacement of joints is given by $w_{st} = p_s - p_t$ where p_s are the joint points for node s given by the linear projection (3.22). A Gaussian scale mixture with component weights π_k encodes relative motion while remaining robust to bursts of large motion,

$$\psi_{st}^{\text{motion}}(x_s, x_t) = \sum_k \pi_k N(w_{st}|u_{st}, V_{st,k}). \tag{3.29}$$

An appearance constancy likelihood captures the similarity of color histograms between parts in consecutive frames. Let $h(x_s) \in \mathbb{R}^J$ be the color histogram over pixels that lie within the part x_s when projected onto the image plane. The appearance likelihood is given by the average element-wise product between histograms,

$$\psi_{st}^{\text{appearance}}(x_s, x_t) = \exp\left(\frac{1}{J} h(x_s)^T h(x_t)\right). \tag{3.30}$$

The hand term ψ_s^{hand} scores the likelihood of hand placement as in Sapp et al. [139] and is only evaluated for the hand region. A linear SVM hand classifier is trained on

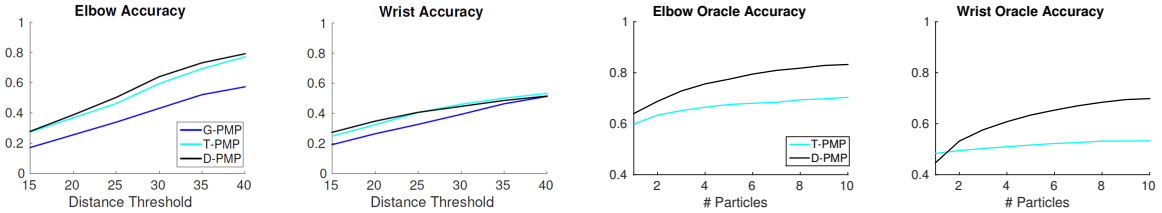


Figure 3.12: **VideoPose2 test results.** *Left:* Elbow and wrist detection accuracy for each PMP algorithm based on a distance threshold ranging from 15 to 40 pixels. *Right:* Detection accuracy of the oracle solution at a 30-pixel detection radius. Particles are chosen in their selection order. Accuracy continues to improve as more D-PMP particles are considered, showing greater diversity in the particle set.

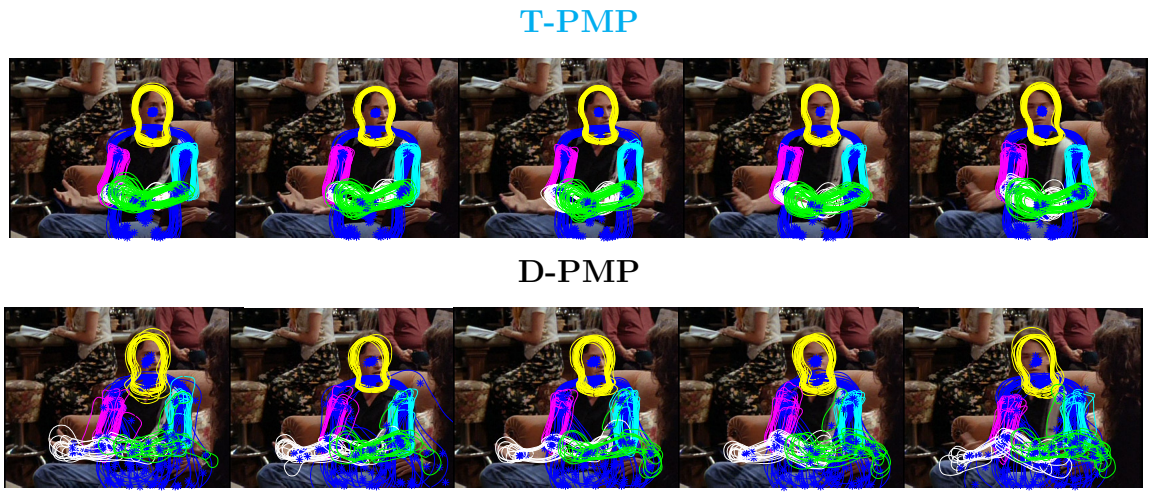


Figure 3.13: **Pose tracking particle diversity.** Final particles for several frames of a VideoPose2 test clip (41). T-PMP (*top*) loses track of the right hand, all particles are concentrated on an alternate mode. D-PMP (*bottom*) maintains greater diversity in the particles, and retains estimates of the right hand at the correct location, and the position preferred by T-PMP.

the flow gradient magnitude, thereby exploiting the notion that hands often exhibit large motion. The hand log-likelihood is the average score of the hand detector over pixels in the projected part hypothesis.

Results

We evaluate tracking accuracy on the VideoPose2 dataset developed by Sapp et al. [139]. The dataset is a benchmark for recent pose tracking publications [188, 43], consisting of 44 clips from the TV shows "Friends" and "Lost", with each clip running 2-3 seconds for a total of 1,286 frames. The dataset is curated to contain frames showing a central figure with upper body visible. We use the standard training-test

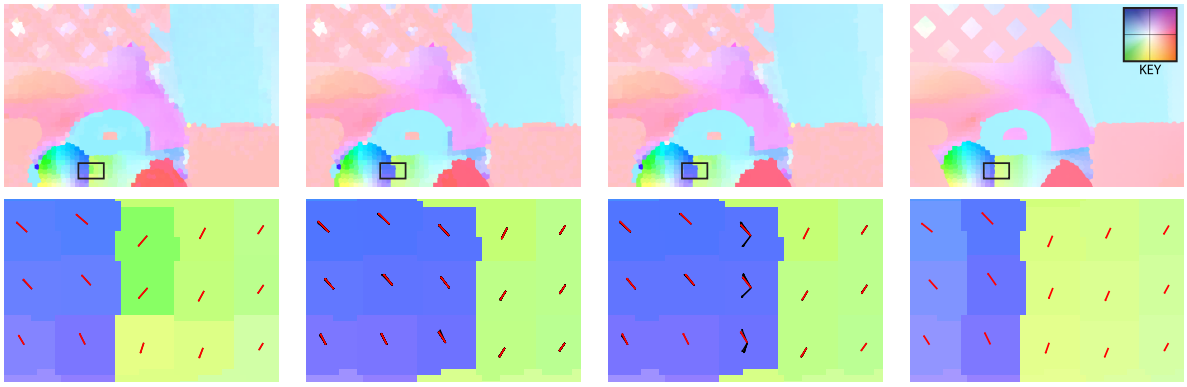
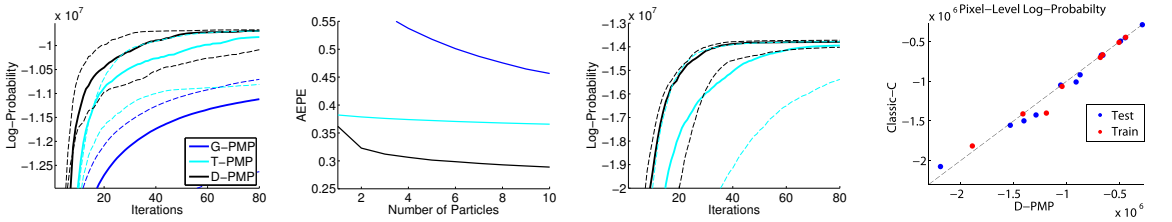


Figure 3.14: **Preserving multiple hypotheses.** *Top Row:* Final flow estimate of each method for the "Rubber Whale" sequence. The color key (top-right) encodes flow vector orientation, color saturation denotes magnitude. *Bottom Row:* Detail of highlighted region showing selected flow particles as vectors (*black*) and the MAP label (*red*). The MAP estimates of D-PMP and T-PMP have higher probability than ground truth, but D-PMP preserves the correct flow in the particle set.

split of 26 training clips and 18 test clips, with one clip recycled for algorithm and model development.

Following [188] we initialize inference using the top scoring solutions given by Flexible Mixture of Parts (FMP) [176] along with uniformly sampled puppets. We use the same DS prior from experiments reported in Section 3.3.1. We find that PMP produces reasonable accuracy in acceptable time using 100 particles and 100 iterations.

For evaluation we compute distance-based detection accuracy of the elbow and wrist; a part is considered *detected* if it falls within a distance threshold ranging from 15 pixels up to 40 pixels. Figure 3.12 (left) reports accuracy of the MAP estimate for each PMP algorithm. D-PMP typically produces the most accurate solution, though T-PMP performs well at large detection thresholds. Estimation failures tend to be the result of poor model fit, as reflected in the oracle accuracy plots, Figure 3.12 (right). Allowing more hypotheses improves the oracle accuracy of D-PMP whereas T-PMP shows little improvement. This suggests that diversity of D-PMP particles preserves high quality estimates that are discarded by T-PMP and not preferred under the model. Figure 3.13 shows final particles of each method, and we see that T-PMP does not keep any accurate right arm particles; by contrast D-PMP maintains hypotheses in both locations.



Log-Probability | Oracle Avg. EPE | Log-Prob. (Test) | Pixel-Level Log-Prob.

Figure 3.15: **Optical flow results.** *Left:* Log-probability quantiles showing median (solid) and best/worst (dashed) MAP estimates versus PMP iteration for 11 random initializations on the Middlebury training set. *Left-Center:* Oracle AEPE over the training set. *Right-Center:* Log-probability quantiles on the test set (G-PMP omitted due to poor performance on training). *Right:* Log-probability of the MAP estimates at the pixel-level model obtained by initializing L-BFGS at the D-PMP solution.

3.3.3 Optical Flow

Given a pair of (grayscale) images I_1 and I_2 in $\mathbb{R}^{M\times N}$ we estimate the motion of each pixel s from one image to the next. This *flow vector* x_s is decomposed into horizontal u and vertical v scalar components. We model flow at the superpixel level using the Classic-C model [159], holding flow constant over each superpixel. Edges are given by the immediate neighbors in I_1.

The pairwise term enforces a smoothness prior on flow vectors via the robust Charbonnier penalty, a differentiable approximation to L_1. This term is approximately quadratic in the range $[-\sigma, \sigma]$ and smoothly transitions to a linear function beyond this range. The potential decomposes additively as $\log \psi_{st} = \phi_{st}^{\text{vert}} + \phi_{st}^{\text{hor}}$ into vertical and horizontal components, defined as follows:

$$\phi_{st}^{\text{hor}}(u_s, u_t) = -\lambda_s \sqrt{\sigma^2 + (u_s - u_t)^2}. \tag{3.31}$$

The spatial smoothness depends on scaling parameter λ_s.

Likelihood potentials $\log \psi_s(x_s) = \phi_s(x_s)$ assume brightness constancy: properly matched pixels should have similar intensities. Each superpixel s contains a number of pixels $\mathcal{I}_s = \{(i_1, j_1), \ldots, (i_k, j_k)\}$, and for each pixel (i, j) we compute the warped coordinates $(\widetilde{i}, \widetilde{j}) = (i + u_s, j + v_s)$. The likelihood penalizes the difference in image intensities, again using the Charbonnier penalty:

$$\phi_s(u_s, v_s) = -\lambda_d \sum_{(i,j)\in\mathcal{I}_s} \sqrt{\sigma^2 + (I_1(i, j) - I_2(\widetilde{i}, \widetilde{j}))^2} \tag{3.32}$$

In computing the warped coordinates we also constrain any pixels which flow outside the image boundary to be exactly on the boundary, $\widetilde{i} = \min(M, \max(0, i + u_s))$. We apply bicubic interpolation for non-integer coordinates.

	Avg. Log-Prob. (p value)	Avg. EPE (p value)
RMP	-2.446E6 (0.008)	1.623 (0.008)
G-PMP	-1.408E6 (0.008)	0.699 (0.008)
T-PMP	-1.212E6 (0.008)	0.382 (0.727)
D-PMP	-1.209E6 (–)	0.362 (–)
Classic-C	–	0.349 (0.727)

Table 3.1: **Optical flow MAP estimates.** Average log-probability and AEPE over 11 random initializations on the Middlebury training set. Reported p values are compared to D-PMP using a Wilcoxon signed rank test, we consider $p < 0.05$ significant.

Results

We evaluate on the Middlebury optical flow benchmark [10] using 11 random initializations. D-PMP and T-PMP utilize 75% neighbor proposals and 25% random walk. We compute SLIC superpixels [1] with region size 5 and regularizer 0.1 resulting in about 5,000 to 15,000 superpixels per image. We use the Charbonnier widths $\sigma = 0.001$ recommended for this model [159], but learn different scaling parameters ($\lambda_s = 16, \lambda_d = 1$) to compensate for our superpixel representation.

Figure 3.15 (left) reports log-probability quantiles over the 8 Middlebury training images. To demonstrate particle diversity we report average endpoint error (AEPE) of the oracle solution in Figure 3.15 (left-center). D-PMP shows a large reduction in AEPE after just a few particles while T-PMP remains nearly flat, suggesting little diversity. In just two dimensions the Gaussian spread of G-PMP particles naturally leads to an error reduction. The benefit of particle diversity is best visualized near object boundaries (see Fig. 3.14).

We also compare to a specialized coarse-to-fine, multiscale inference algorithm for Classic-C[1], using the recommended settings and with the median filter disabled. We also compare to RMP on a fixed regular discretization of 200 flow vectors. As shown in Table 3.1, D-PMP yields significantly higher probability solutions, but is equivalent to T-PMP in AEPE. D-PMP also achieves equivalent results to Classic-C optimization, which is highly tuned to the Middlebury dataset.

We cannot directly compare probability of the Classic-C and D-PMP solutions, because the former models flow at the pixel level. Instead, using L-BFGS initialized from the D-PMP solution, we optimize the pixel level model and compare log-probability of the result with Classic-C for both training and test sequences (Fig. 3.15 (right)). Again, even compared to a highly-tuned specialized optimization method, D-PMP achieves statistically equivalent results.

[1] `http://people.seas.harvard.edu/~dqsun`
Experiments use code accessed on 06 February 2015.

3.4 Discussion

Particle max-product methods provide a class of general-purpose MAP inference algorithms for high-dimensional, continuous graphical models. These methods balance the efficiency of variational approximations with the flexibility of particle-based approximations. Particle selection, necessary to control computation, induces degeneracies reminiscent of sequential Monte Carlo when done naively. To avoid such degeneracies we introduce a diverse particle selection approach which preserves multiple local modes.

Through this diversity our algorithm (D-PMP) is better able to reason globally about competing hypotheses, and is more robust to initialization. With D-PMP we obtain accurate pose estimates on a challenging dataset, and for images with multiple people we preserve hypotheses on each figure even without an explicit multi-person model. Motivated by results on single images we explore pose tracking in video where temporal dynamics induce cyclic dependencies in the underlying MRF. To support loopy models we use a reweighted variant of max-product message passing and reformulate diverse particle selection to incorporate the RMP spanning tree distribution. When D-PMP is applied to pose estimation in video we obtain superior accuracy to competing particle max-product methods.

While pose estimation serves as a motivating application we stress that D-PMP is not specialized to this task, but is a general inference algorithm. We demonstrate the generality of our approach for a very loopy optical flow model. Using the same implementation of D-PMP we obtain competitive performance with standard inference algorithms for this task. A MATLAB library, built on UGM [140], implementing these methods is available[2].

[2] `http://www.cs.brown.edu/~pachecoj`

Chapter 4

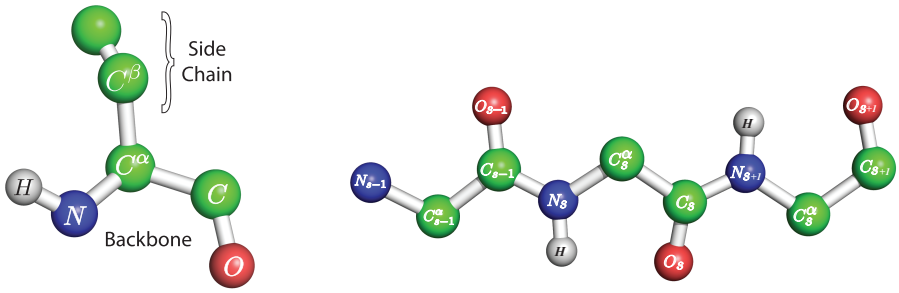
Protein Structure Prediction

One of the most important tasks in computational biology is the ability to predict the 3D structure of a protein from its amino acid sequence. The structure is used to locate drug binding sites, determine biological function, and understand diseases linked to misfolding. While protein folding is an ambitious goal several important sub-problems in structure prediction are more attainable, such as side chain prediction, protein docking, binding, and design. We focus on side chain prediction and discuss how D-PMP inference can advance the state-of-the-art for this problem.

The majority of existing approaches to protein structure prediction rely on a fixed discretization of the structural representation. Optimization over this discrete space is generally performed by simulated annealing Monte Carlo [137], discrete search [58], or by max-product belief propagation [59]. We show that optimization over the continuous space of side chain placement is able to capture structural details that discrete methods cannot.

Several related methods consider continuous optimization of protein structure. While they do not consider side chain prediction Peng et al. [128] apply the greedy PMP variant (Ch. 3) for protein folding. Ghorie et al. apply the Metropolis variant of PMP [95] lacking the diverse particle selection that we propose here [65]. Such diverse configurations are particularly important in protein structure since proteins ar known to take multiple stable configurations, depending on function [165, 100, 109].

Most structure prediction methods seek several distinct configurations via independent MCMC chains [59, 58, 137]. With few exceptions [57] little attention has been paid to finding a diverse set of conformations. Lang et al. [100] propose a method to discover multiple distinct side chain configurations by drawing samples from an electron density map, however this approach requires experimental X-ray data. Through our diverse particle selection we show that D-PMP is capable of preserving solutions at multiple local optima.



Amino Acid **Protein Backbone**

Figure 4.1: Amino acid (left) consisting of a backbone and side chain molecule, with individual atoms labeled. Amino acids bond to form a repeated protein backbone (right).

4.1 Side Chain Prediction

Each of the twenty biologically relevant amino acids consist of a central carbon atom (C^α) surrounded by an amine group (NH_2) and a carboxylic acid ($COOH$). Neighboring amino acids bond to form a repeated backbone structure, shown in Figure 4.1. In addition to the backbone, a molecule unique to each amino acid type, known as a *side chain*, is attached to C^α. Assuming a known protein backbone configuration, side chain prediction attempts to estimate the configuraiton of side chain molecules along the protein chain.

4.1.1 Amino Acid Side Chains

A compact angular representation of side chains is given by *dihedral angles* (Fig. 4.2 (left)), which encode the relative rotation of intersecting planes. Unlike an encoding of atomic coordinates, which allows physically impossible structures, an angular representation exploits structural constraints of rigidly bonded atoms, thereby reducing the total degrees of freedom. The coordinates of each atom are recovered from an angular encoding by assuming ideal bond geometry [49].

Side chain composition varies by amino acid type with up to four dihedral angles $\chi_1, \ldots, \chi_4$ for each amino acid type. Each dihedral angle describes a rotation about neighboring carbon atom bonds, for example χ_1 is a rotation about the backbone C^α and the first side chain carbon C^β (Fig. 4.2 (right)).

4.1.2 Discrete Rotamer Optimization

While side chain angles may take any configuration in the continuum $\chi_i \in [0°, 360°)$ they tend to cluster around one of three stable configurations in observed X-ray structures $\chi_i \in \{60°, 180°, 300°\}$. These configurations, known as *rotamers* [25],

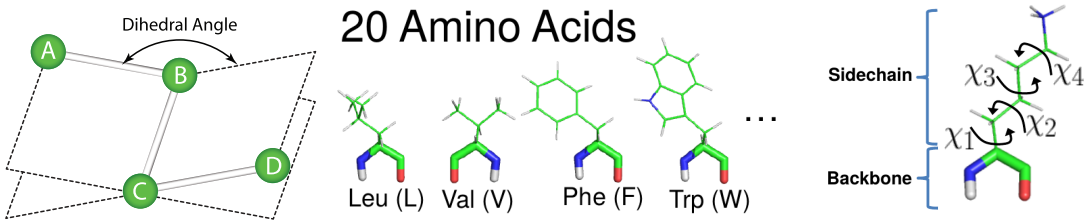
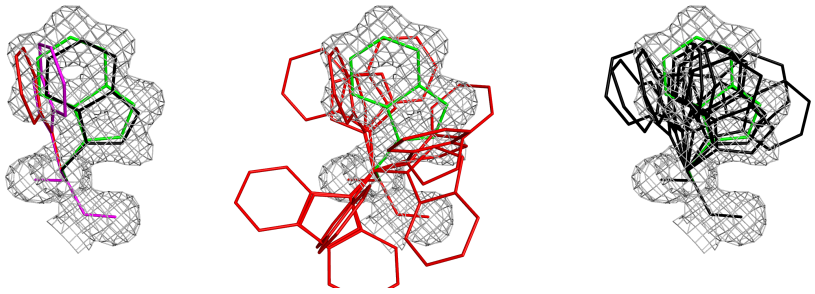


Figure 4.2: Dihedral angles (left) define relative rotation between intersecting planes defined by bonded atoms. Among the 20 amino acid types (center) the number of side chain dihedral angles varies (right).



Estimate Rotamers D-PMP Particles

Figure 4.3: **Non-rotameric side chains.** *Left:* X-ray (green), RMP (red), Rosetta (magenta) and D-PMP (black) estimates. *Center:* Standard rotamers are all poor approximations. *Right:* Final D-PMP particles all overlapping the level set of the electron density (mesh). (PDB: 1GK9, Trp154) [146]

are not uniformly preferred, but instead occur with varying marginal probability in observed X-Ray structures. *Rotamer libraries* encode these marginal probabilities while marginalizing [130, 108] or conditioning [47, 46] on backbone configurations.

Methods for side chain prediction rely almost exclusively on discrete optimization over rotamer configurations, to avoid minimizing a continuous energy function. The optimization is NP-hard [149], but can be solved in some cases using dead-end elimination (DEE) [42, 68], a branch-and-bound method for reducing the size of the search space. Large proteins, however, may remain intractable and search methods based on A^* [104, 58] or simulated annealing [105, 76, 97] are used to optimize the energy. Heuristic techniques have also been developed, such as the widely used SCWRL [29] and TreePack [174] algorithms.

Max-product BP is effective for estimating side chains from the rotamer discretization. Fromer et al. apply standard max-product (MP) to side chain prediction and protein design [58]. Yanover et al. investigate standard MP and reweighted max-product (RMP) compared to a general-purpose LP solver (CPLEX) for side chain prediction [178, 179]. When MP does converge it typically produces the best solution in the shortest time, and RMP almost always finds the solution significantly faster than CPLEX. Weiss et al. look at the more general class of convex BP algorithms,

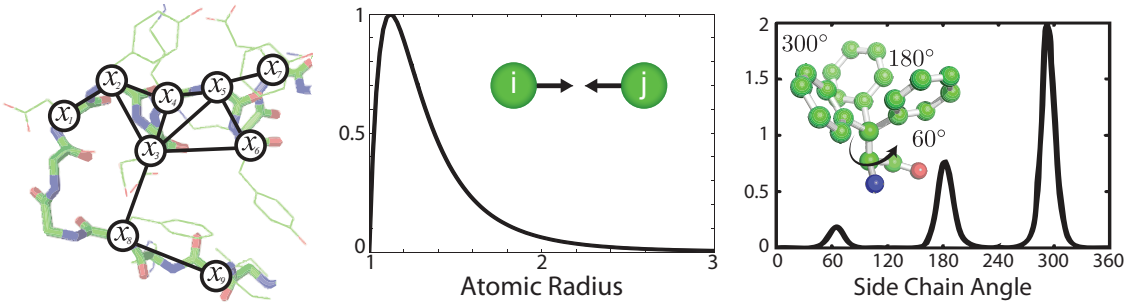


Figure 4.4: **Protein structure energetics.** *Left:* Pairwise MRF derived by adding an edge between each pair of residues within a 10 Å radius. *Center:* The Lennard-Jones 6-12 potential between two atoms is extremely repulsive at close range and attractive at moderate ranges. *Right:* The Dunbrack rotamer probability for a single side chain angle is a Gaussian mixture to the marginal statistics of observed side chain dihedral angles.

of which reweighted BP is one instance [171], where they focus on conditions under which the MAP can be provably obtained. Finally, Sontag et al. explore provably convergent MPLP optimization for side chain packing [153].

4.2 Continuous Side Chain Optimization

Rotamer libraries enable efficient optimization, but they assume an ideal geometry known to be violated by some proteins [162, 146]. Moreover, the rotamer discretization is a coarse approximation that fails to capture fine details of side chain placement. By applying particle max-product we instead perform side chain prediction by minimizing a continuous energy function that is motivated by well-supported thermodynamic interpretations [9, 145, 54, 28, 109].

4.2.1 Graphical Model of Side Chain Placement

We propose a simple model of side chain prediction as a pairwise MRF. Nodes $s \in \mathcal{V}$ represent amino acids and edges $(s,t) \in \mathcal{E}$ join nearby amino acids with C^α atoms in a 10 Å radius,

$$p(x) \propto \prod_{s \in \mathcal{V}} \psi_s^{\text{rot}}(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}^{\text{LJ}}(x_s, x_t). \tag{4.1}$$

Our approach is similar to the discrete formulation of Yanover et al. [179], but defined over a continuous latent state. For a side chain with D dihedral angles the latent state is a vector of angles $x_s \in [0°, 360°)^D$. Pairwise terms model the physical interaction between nearby amino acids and unary terms encode the likelihood of side chain configurations.

The repulsive and attractive forces of atomic interaction are described by the *van der Waals* forces. The *Lennard-Jones* potential [87] gives a numerical approximation to these forces between for a pair of amino acids (s, t),

$$\log \psi_{st}^{\mathrm{LJ}}(x_s, x_t) = \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} 4\epsilon \left[\left(\frac{\sigma}{r_{ij}(x_s, x_t)} \right)^6 - \left(\frac{\sigma}{r_{ij}(x_s, x_t)} \right)^{12} \right], \tag{4.2}$$

where $r_{ij}(x_s, x_t)$ is the relative distance between a pair of atoms, N_s is the number of atoms in the s^{th} amino acid, ϵ controls the strength of attraction, and σ is the cutoff distance where atoms do not interact.

The likelihood of proposed structural properties encodes the statistics of observed X-ray structures. This *rotamer probability* scores the likelihood of side chain dihedral angles based on the rotamer library of Dunbrack et al. [47, 46]. The rotamer probability is given by a Gaussian mixture over M rotamers,

$$\psi_s^{\mathrm{rot}}(x_s) = \sum_{m}^{M} \pi_m N(\chi_s | \mu_m, \Sigma_m). \tag{4.3}$$

The mean μ and covariance Σ of each mixture component are implicitly conditioned on the backbone dihedral angles at each amino acid. Backbone angles are binned in 20° increments to produce a finite set of moments.

4.2.2 Resolving Ties in the Conformation

In PMP we resolve ties using an approach similar to one proposed for discrete MRFs [171]. Recall that, given RMP fixed point messages m_{us} and edge appearance probabilities ρ_{us}, pseudo-max-marginals for each node are (Sec. 2.3.4):

$$\nu_s(x_s) \propto \psi_s(x_s) \prod_{u \in \Gamma(s)} m_{us}(x_s)^{\rho_{us}}.$$

A similar definition holds for pairwise pseudo-max-marginals ν_{st} on each edge $(s, t) \in \mathcal{E}$. These pseudo-max-marginals admit a provably MAP solution x^* if a consistent labeling exists in the set of maxima [168]:

$$x_s^* \in \arg\max_{x_s} \ \nu_s(x_s), \quad (x_s^*, x_t^*) \in \arg\max_{x_s, x_t} \ \nu_{st}(x_s, x_t).$$

For continuous distributions exact ties rarely exist, but small numerical errors in the estimated pseudo-max-marginals can perturb the particle that is inferred to be most likely, and lead to joint states with low probability due to conflicted edges. This

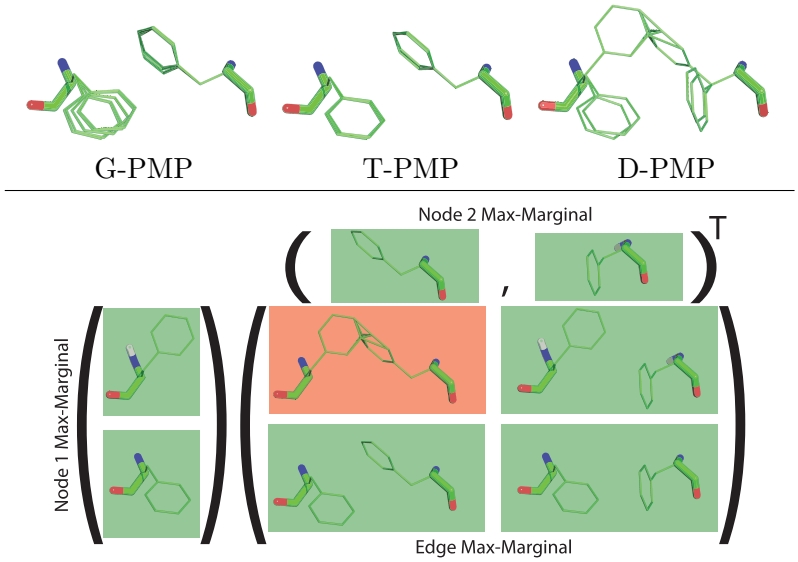


Figure 4.5: **Label Conflicts.** *Above:* Selected side chain particles of two amino acids (PDB: 1QOW). Diversity in the D-PMP particle set presents more opportunity for an inconsistent labeling. *Below:* Naively maximizing the node max-marginal over two tied states can produce a very unlikely joint configuration.

problem is common in the side chain model and as illustrated in Fig. 4.5, the diversity in the D-PMP particles makes conflicts more likely.

To address ties we relax the set of optima to those states with pseudo-max-marginal values within ϵ of the maximum:

$$\mathrm{OPT}(\nu_s) = \{x_s^* : |\nu_s(x_s^*) - \arg\max_{x_s} \nu_s(x_s)| \leq \epsilon\}. \tag{4.4}$$

Let $\mathcal{V}_T$ be the set of tied nodes with more than one near-maximal state and $\mathcal{E}_T = \mathcal{E} \cap (\mathcal{V}_T \otimes \mathcal{V}_T)$ the edges joining them. Let x_{NT}^* be the unique assignments for non-tied nodes. Construct an MRF over the remaining tied nodes as

$$p_T(x_T) \propto \prod_{s \in \mathcal{V}_T} \widetilde{\psi}_s(x_s) \prod_{(s,t) \in \mathcal{E}_T} \psi_{st}(x_s, x_t), \tag{4.5}$$

with the conditioned node potentials

$$\widetilde{\psi}_s(x_s) = \psi_s(x_s) \prod_{t \in \Gamma(s) \setminus \mathcal{V}_T} \psi_{st}(x_s, x_t^*). \tag{4.6}$$

We label the remaining nodes $x_T^* = \arg\max_{x_T} p_T(x_T)$ using the junction tree algorithm. If the junction tree contains a unique maximizer, then $x^* = (x_T^*, x_{NT}^*)$ is the global MAP over the particles $\mathbb{X}$. This guarantee follows from the reparameterization property of pseudo-max-marginals and Theorem 2 of [171]. Clique size is reduced by eliminating non-tied nodes, and by constraining labels to the set of tied states $x_T \in \mathrm{OPT}(\nu_s)$.

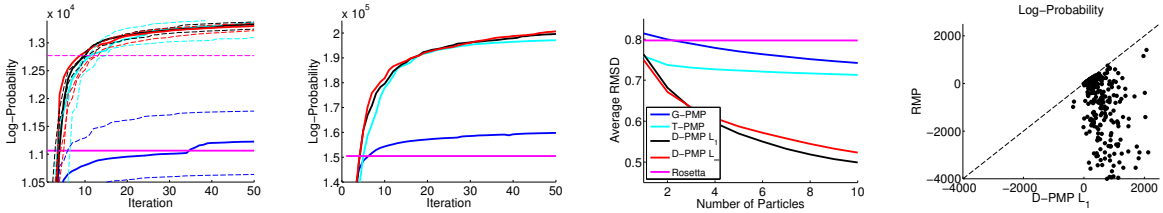


Figure 4.6: **Side chain prediction.** We compare each method and both L_1 and L_∞ diverse selection methods. *Left:* Total log-probability over 20 proteins. Median (solid) and best/worst (dashed) results on 11 random initializations. *Left-Center:* Total log-probability for 370 proteins. *Right-Center:* RMSD (in angstroms Å) of the oracle solution on larger set. *Right:* Log-probability of all 370 proteins versus the fixed rotamer discretization with RMP inference.

4.3 Experimental Results

In the following experiments we explore PMP for protein side chain prediction. We show that our approach to diverse particle selection captures multiple distinct solutions which encode the conformational diversity that proteins are known to exhibit [165, 100, 109]. For evaluation of the side chain energy (4.1) we use the Rosetta molecular modeling package [137].

In addition to evaluating energy potentials Rosetta is also our main comparison benchmark. Rosetta predicts structure using a *Monte Carlo plus minimization* approach proposed by Li and Scheraga [107]. Proposal distributions vary depending on whether the goal is side chain prediction or folding. For side chain prediction Metropolis proposals are sampled from the rotamer library, followed by continuous quasi-Newton optimization [137].

In Figure 4.6 we compare estimation accuracy of side chain placement for each algorithm on two sets of proteins selected from the Protein Data Bank[1]. Particle max-product is configured with 50 particles for 50 iterations. D-PMP and T-PMP proposals are 50% random walks from Gaussians wrapped to account for angular discontinuities, and 50% samples from the rotamer marginals. Neighbor-based proposals are not used, due to the complex transformation between dihedral angles and atomic coordinates.

D-PMP typical runs outperform competitors. We run each method from 11 random initialization on a small set (20 proteins) and report quantiles of total log-probability (Fig. 4.6 (left)). Both D-PMP and T-PMP outperform G-PMP, due to their ability to exploit the model likelihood through rotamer proposals, with D-PMP showing the tightest confidence intervals. The second set is larger (370 proteins) and

[1] http://www.pdb.org

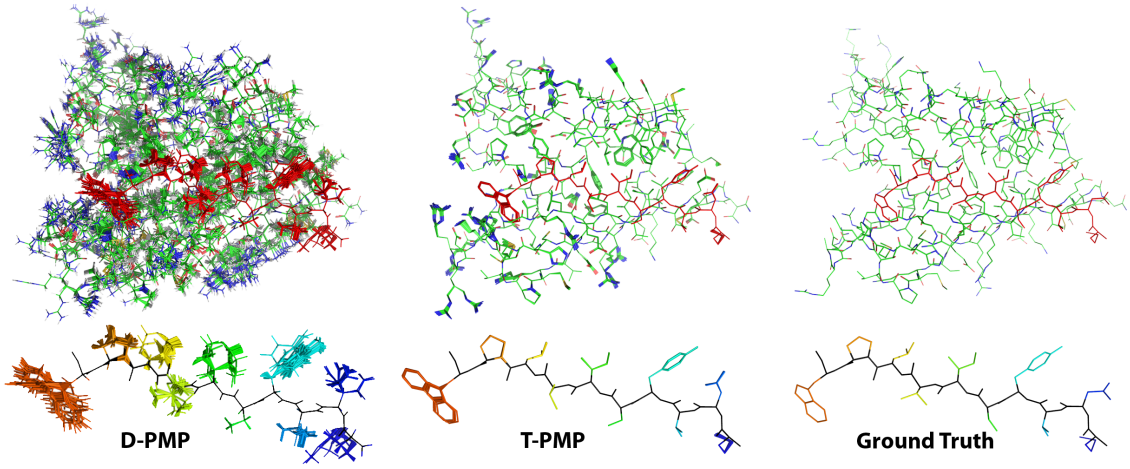


Figure 4.7: **Side chain particles.** *Top Row:* Final particles for T-PMP and D-PMP, and the ground truth conformation of a single protein (PDB ID: 2H8E). Region marked in red is detailed below. *Bottom Row:* Closeup of first ten amino acids, showing the fixed backbone (black) and final particles colored by backbone location. D-PMP preserves more diverse particles in areas of uncertainty.

we report the total log-probability of a single run for each method (Fig. 4.6 (left-center)).

D-PMP preserves a diverse configuration of side chains. Fig. 4.7 shows a qualitative comparison of diversity between D-PMP and T-PMP for a single protein. To measure diversity we report RMSD of the oracle solution (Fig. 4.6); D-PMP shows a substantial improvement in accuracy after only a few particles. We also compare the submodular particle selection (L_1) with the *minimax* formulation (L_∞); both preserve diversity similarly, but the former offers stronger theoretical justification.

4.4 Discussion

The results in this chapter further demonstrate effectiveness of D-PMP inference for arbitrary continuous MRFs. Indeed, the side chain model explored here bears little resemblance to the models of human pose estimation which were the focus of Chapter 3. One significant difference is the Lennard-Jones potential, which describes compatibility of nearby atoms. This energy is extremely peaked at close ranges leading to frequent ties in the pseudo-max-marginals, not observed in previous applications. We show that resolving these ties is crucial to avoid extreme penalties in the MAP labeling, and can be done in a way previously proposed for discrete MRFs.

Through our departure from the standard approach to side chain prediction, based

on discrete energy minimization, we show that optimization of a continuous energy function avoids several inaccuracies imposed by the rotamer discretization. First, side chains tend to cluster around the rotamer configurations, but significant variation exists around these modes which is not captured by a discretization [18]. Secondly, the final dihedral angle of many residues often does not obey the rotamer discretization leading to continuous density estimates [147], a phenomenon that is naturally handled by optimizing continuous energies. Finally, experimental results show that some side chains disobey the rotamer configurations entirely, or can be found in multiple configurations [146], both cases are supported by D-PMP (Fig. 4.3).

Chapter 5

Variational Inference for Generalized Gaussian Mixtures

The BP message updates (2.25) are only tractable for discrete and Gaussian probability models, and we begin this chapter with an exploration of inference algorithms on a broader class of models. In particular we develop EP inference for a broad class of target tracking models in the presence of unwanted *clutter* detections modeled by Gaussian mixture emission probabilities (Sec. 5.1). Our approach unifies a number of classical techniques in probabilistic target tracking into a common framework, and also generalizes to yield new algorithms not previously considered in the literature.

Going further, we introduce a new class of inference algorithms (Sec. 5.2) based on applying Lagrange multiplier methods to directly minimize a variational objective function. Stationary points of this objective, known as the *Bethe free energy*, correspond to fixed points of LBP and EP. By this correspondence the resulting algorithms share the same solution set as their message passing counterparts, but are provably convergent and exhibit superior stability properties by avoiding degenerate conditions which often occur in Gaussian LBP and EP for continuous models.

5.1 Robust Target Tracking in Clutter

Within the Bayesian approach to target tracking there are two competing models for representing uncertainty in observation assignments. The *probabilistic data association filter* (PDAF) [11] assumes at most one true target detection per time step and incorporates observations sequentially via a single forward pass. The *probabilistic multi-hypothesis tracker* (PMHT) [155] allows for an arbitrary number of target detections and adapts the *expectation maximization* (EM) algorithm to iteratively estimate smoothed state updates from a fixed batch of data. Both of these approaches

approximate the intractable posterior over the target state with a single Gaussian.

In contrast to previous approaches, which consider only a single observation assignment model, we derive EP inference for both models. To differentiate the algorithms from their probability models we refer to the *dependent* observation assignment model underlying the PDAF, and the *independent* observation assignment model underlying the PMHT. We also consider single Gaussian and Gaussian mixture approximations of the posterior over target state. In the next sections we briefly outline each of the algorithms. For a more detailed discussion see the published report [124].

5.1.1 Expectation Propagation for Target Tracking

We begin by briefly reviewing the EP updates first discussed in Sec. 2.3.2. Consider a joint distribution which factorizes according to

$$p(x \mid \mathcal{D}) \propto p_0(x) \prod_i \psi_i(x) \tag{5.1}$$

with latent variables x, prior distribution $p_0(x)$, and observed data $\mathcal{D}$ encoded via non-negative factors $\psi_i(x)$. We choose an approximating distribution $q(x)$ that is in a tractable *exponential family* [169] of distributions, with matched factorization

$$q(x) = p_0(x) \prod_i \widetilde{\psi}_i(x) \approx p(x \mid \mathcal{D}). \tag{5.2}$$

We refer to $\widetilde{\psi}_i(x)$ as *messages*, which can be thought of as local approximations. EP provides a means for iteratively refining each $\widetilde{\psi}_i(x)$ such that $q(x)$ approximates the true posterior $p(x \mid \mathcal{D})$. At each iteration, EP updates the posterior and factor approximations according to the following procedure:

$$\begin{aligned}
q^{\setminus i}(x) &= q(x)/\widetilde{\psi}_i(x) && \text{(Cavity)} \\
\hat{p}(x) &\propto q^{\setminus i}(x)\psi_i(x) && \text{(Augmented Distribution)} \\
q^{\text{new}}(x) &= \arg\min_q \; \mathrm{D}(\hat{p}(x) \parallel q(x)) && \text{(KL Projection)} \\
\widetilde{\psi}_i(x) &\propto q^{\text{new}}(x)/q^{\setminus i}(x) && \text{(New Message)}
\end{aligned}$$

The KL projection can be computed in closed form via moment-matching (c.f. Sec. 2.2). The messages $\widetilde{\psi}_i(x)$, as well as the cavity distribution $q_{\setminus i}(x)$, are members of an *unnormalized* exponential family. EP does nothing to explicitly enforce their normalizability– an issue we directly address in Sec. 5.2. For the KL projection to be well-posed, the augmented distribution $\hat{p}(x)$ must be normalizable. If it is not then we "halt" the update, leaving the message unchanged. For more details on EP in general, see [114, 169].

In the following sections we derive EP inference algorithms for the *dependent* clutter assignment model underlying the PDAF, and the *independent* clutter assignment model that PMHT assumes. We consider approximations of state distributions by two exponential families, single Gaussian and Gaussian mixture. We do not apply the mixture approximation to the independent assignment model, where the mixture size is exponential in the number of observations per timestep.

For all of the target tracking models we consider the joint distribution factorizes as a time series with first-order Markov dependence:

$$p(X, Z) = \frac{1}{Z} p_0(x_0) \prod_{t=1}^{T} \psi_t(x_{t-1}, x_t) \varphi_t(x_t, z_t) \tag{5.3}$$

where the target state at scan t is $x_t \in \mathbb{R}^n$ with prior $p_0(x_0) = N(x \mid \mu_0, V_0)$ and linear Gaussian target dynamics $\psi_t(x_{t-1}, x_t) = N(x_t \mid F x_{t-1}, Q)$ where $F, Q \in \mathbb{R}^{n \times n}$. The observation likelihoods $\varphi_t(x_t, z_t)$ encode the assignment model, and depend implicitly on observed data $y_t = \{y_t^i\}_{i=1}^{M_t}$.

Clutter Assignment Models

Under the dependent assignment model, at most one detection per timestep is related to the target state. Assignments are encoded as $z_t \in \{0, \ldots, M_t\}$, where $z_t = 0$ indicates that all observations are clutter. Otherwise, $y_t^{z_t}$ is target-generated:

$$\varphi_t^D(x_t, z_t) = \delta_{z_t,0} \lambda_0 + \sum_{i=1}^{M_t} \delta_{z_t,i} \lambda_i N(y_t^i \mid H x_t, R) \tag{5.4}$$

Here, $\delta_{\cdot,\cdot}$ is the Kronecker delta, $H \in \mathbb{R}^{m \times n}$ and $R \in \mathbb{R}^{m \times m}$. The overall potential is a mixture of M_t Gaussians and a constant. Typically, $\lambda_0 = (1 - P_d) N(y_t^i \mid 0, \Sigma_0)$ and $\lambda_i = \frac{P_d}{M_t} p(z_t = i)$ for some probability of detection P_d.

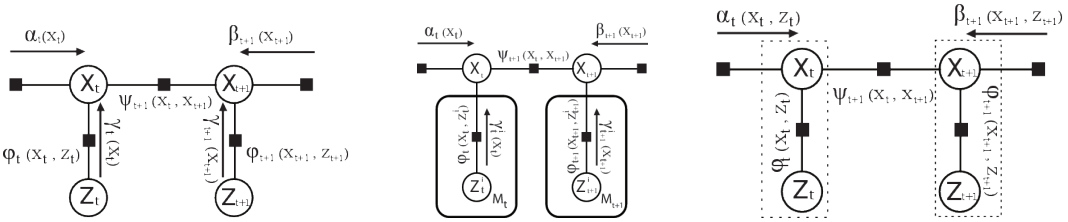
The independent assignment model assumes the M_t detections are marginally independent, where $z_t^i = 0$ if detection i is clutter, and $z_t^i = 1$ if it is related to the target:

$$\varphi_t^I(x_t, z_t^i) = \delta_{z_t^i,0} \lambda_0 + \delta_{z_t^i,1} \lambda_1 N(y_t^i \mid H x_t, R) \tag{5.5}$$

The overall observation likelihood at time t is then the product $\prod_{i=1}^{M_t} \varphi_t^I(x_t, z_t^i)$, a mixture of $\mathcal{O}(2^{M_t})$ Gaussians plus a constant term.

EPD: Dependent Assignment, Single Gaussian

We begin with a Gaussian marginal posterior approximation $q_t(x_t) = N(x_t \mid m_t, V_t)$ defined as the product of a *forward* message $\alpha_t(x_t)$, a *backward* message $\beta_t(x_t)$, and



(a) EPD (b) EPI (c) EPD+

Figure 5.1: Factor graphs illustrating the joint factorization and messages underlying three EP tracking algorithms. EPD: Dependent observation assignments, single Gaussian state distributions. EPI: Independent observation assignments, single Gaussian state distributions. EPD+: Dependent observation assignments, Gaussian mixture state distributions.

a *measurement* message $\gamma_t(x_t)$:

$$q_t(x_t) \propto \alpha_t(x_t)\gamma_t(x_t)\beta_t(x_t) \approx p(x_t \mid Y_1^T) \tag{5.6}$$

The messages are parameterized as unnormalized Gaussians in information form,

$$\alpha_t(x_t) = s_t^\alpha \exp(-\frac{1}{2}x_t^T \Lambda_t^\alpha x_t + x_t^T \eta_t^\alpha), \tag{5.7}$$

with similar definitions for $\beta_t(x_t)$ and $\gamma_t(x_t)$. Figure 5.1(a) shows a factor graph [169] for this model along with overlays denoting the direction and type of messages. The forward pass augmented distribution at scan t yields a Gaussian density,

$$\hat{p}_t(x_t) \propto \gamma_t(x_t)\beta_t(x_t) \int_{\mathcal{X}_{t-1}} \alpha_{t-1}(x_{t-1})\, \gamma_{t-1}(x_{t-1})\, \psi_t(x_{t-1}, x_t)\, dx_{t-1} \tag{5.8}$$

The EP projection step introduces no approximation, so $q_t^{\text{new}} = \hat{p}_t$. The forward messages $\alpha_t(x_t)$ are as in a conventional Kalman filter, and the reverse messages $\beta_t(x_t)$ as in a two-pass Kalman smoother. In contrast, the measurement messages $\gamma_t(x_t)$ involve a projection step since the augmented distribution is non-Gaussian:

$$\hat{p}_t(x_t) \propto \sum_{z_t=0}^{M_t} \alpha_t(x_t)\varphi_t^D(x_t, z_t)\beta_t(x_t) \tag{5.9}$$

The projection $q_t^{\text{new}} \propto \arg\min_q \mathrm{D}(\hat{p}_t \| q)$ matches the mean and variance of the Gaussian mixture $\hat{p}_t(x_t)$. The measurement message update is $\gamma_t^{\text{new}}(x_t) = \frac{q_t^{\text{new}}(x_t)}{\alpha_t(x_t)\beta_t(x_t)}$.

A single forward pass of this algorithm, iteratively updating α_t and γ_t, is equivalent to the PDAF [11]. To see this, note the correspondence between the PDAF prediction step and the calculation of the forward messages α_t. Similarly, the PDAF association probabilities correspond to the mixture weights of the augmented distribution of Eq. (5.9). The projection step yields a Gaussian posterior $q_t(x_t)$, the mean of which corresponds to the minimum mean square error (MMSE) state prediction of PDAF.

Further iterations of EPD provide a novel way of generalizing PDAF to produce smoothed state estimates. Each iteration has linear complexity $\mathcal{O}(N)$, where $N = \sum_{t=1}^{T} M_t$ is the total number of detections.

EPI: Independent Assignment, Single Gaussian

As in the EPD algorithm, we approximate the state posterior via a single Gaussian distribution:

$$q_t(x_t) \propto \alpha_t(x_t) \left(\prod_{i=1}^{M_t} \gamma_t^i(x_t) \right) \beta_t(x_t) \approx p(x_t \mid Y_1^T) \tag{5.10}$$

Note that we have a separate measurement message $\gamma_t^i(x_t)$ for each observation, and the posterior depends on the product of all of these messages. Figure 5.1(b) shows a factor graph for this model with overlays indicating the forward, backward, and measurement messages. The forward pass augmented distribution at scan t yields a Gaussian density,

$$\hat{p}_t(x_t) \propto \beta_t(x_t) \prod_{i=1}^{M_t} \gamma_t^i(x_t) \int_{\mathcal{X}_{t-1}} \alpha_{t-1}(x_{t-1}) \prod_{i=1}^{M_{t-1}} \gamma_{t-1}^i(x_{t-1}) \psi_t(x_{t-1}, x_t) \, dx_{t-1}$$

This is Gaussian, so as in EPD the forward and backward messages correspond to conventional Kalman filters and smoothers. The measurement message update at each scan is equivalent to an instance of EP for the *clutter problem* [114].

One full iteration of EPI has linear complexity $\mathcal{O}(N)$, where N is again the total number of detections. This algorithm does not appear to be equivalent to classical tracking algorithms, for any message schedule. EPI assumes the same assignment model as the PMHT, but the inference algorithms are different.

EPD+: Dependent Assignment, Gaussian Mixture

Returning to the dependent assignment model of EPD, we extend EP to employ a more flexible, Gaussian mixture marginal approximation. A closely related algorithm has been used for inference in switching state-space models [73]. Tractability of the posterior is maintained by limiting the marginal at scan t to a mixture approximation with M_t modes, one for each possible clutter assignment z_t,

$$q_t(x_t, z_t) = \sum_{j=0}^{M_t} \delta_{z_t,j} p_{t,j} N(x_t \mid m_{t,j}, V_{t,j}) \approx p(x_t, z_t \mid Y_1^T)$$

Note that unlike the simpler EPD approximation, $q_t(x_t, z_t)$ is defined over the target state x_t and assignments z_t jointly. Measurement messages are not necessary, because

the measurement potential lies in the approximating family. We define the marginal as the product of forward and backward messages $q_t(x_t, z_t) \propto \alpha_t(x_t, z_t)\beta_t(x_t, z_t)$. The messages are parameterized as unnormalized Gaussian mixtures,

$$\alpha_t(x_t, z_t) = \sum_{j=0}^{M_t} \delta_{z_t,j} p^{\alpha}_{t,j} \exp(-\frac{1}{2} x_t^T \Lambda^{\alpha}_{t,j} x_t + x_t^T \eta^{\alpha}_{t,j}) \tag{5.11}$$

with a similar definition for $\beta_t(x_t, z_t)$. Figure 5.1(c) depicts a factor graph representation of this model with overlays for the forward and backward messages. The augmented distribution in the forward pass at scan t is:

$$\hat{p}_t(x_t, k) \propto \beta_t(x_t, k)\varphi^D_t(x_t, k) \sum_{j=0}^{M_{t-1}} \int_{\mathcal{X}_{t-1}} \alpha_{t-1}(x_{t-1}, j)\psi(x_{t-1}, x_t)\, dx_{t-1}.$$

For each candidate $z_t = k$, the augmented distribution $\hat{p}_t(x_t, k)$ is a Gaussian mixture with M_{t-1} components. We project each of these mixtures to a single Gaussian $q^{\text{new}}_t(x_t, z_t = k)$ with matched mean and covariance. The posterior approximation q^{new}_t is then a mixture of $M_t + 1$ Gaussians, indexed by z_t.

The updates of backward messages $\beta_t(x_t, z_t)$ proceed similarly to the forward pass. A single forward pass of EPD+ is similar to the *Gaussian Pseudo-Bayesian estimator of second order* (GPB2) [11], which is a forward filter for estimation in a *switching linear dynamical system* (SLDS). One or more forward and backward passes of EPD+ correspond to smoothed generalizations of GPB2, and thus novel algorithms for tracking in clutter. If there are M detections per time step, one iteration of EPD+ has computational complexity $\mathcal{O}(TM^2) = \mathcal{O}(NM)$. This is greater than EPD but still linear in T, and often tractable.

KNN: Nearest Neighbor Association Baseline

The Kalman filter with nearest neighbor association (KNN) provides a baseline comparison [5]. Given approximate filtered marginals $\hat{p}_t(x_t) = N(x_t \mid m_t, P_t)$, we predict the state evolution as follows:

$$\hat{p}(x_{t+1} \mid Y_1^t) = N(x_{t+1} \mid Fm_t, Q + FP_tF^T)$$

We refer to $\hat{x}_{t+1|t} = Fm_t$ as the predicted target state and $\hat{P}_{t+1|t} = Q + FP_{t-1}F^T$ as the predicted covariance. The predicted measurement is given by $\hat{y}_{t+1} = H\hat{x}_{t+1|t}$. Assuming Gaussian noise, the most likely associated measurement can be selected based on the detection nearest to $\hat{y}_{t+1}$:

$$\hat{z}_{t+1} = \underset{z \in \{1,\ldots,M_t\}}{\arg\min} \|\hat{y}_{t+1} - y^z_{t+1}\|_2^2$$

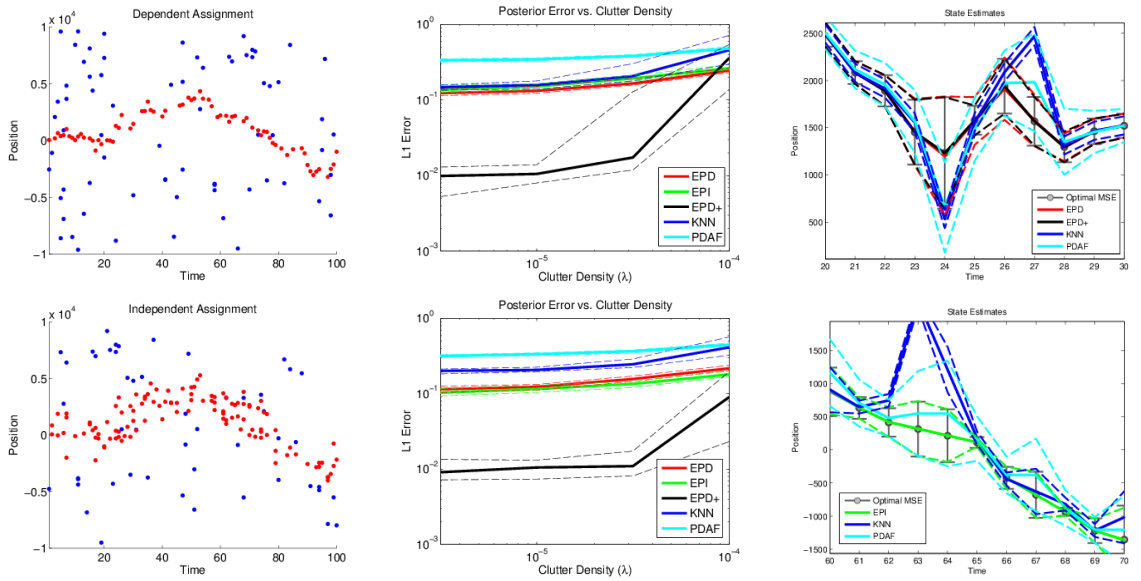


Figure 5.2: Data sampled from the dependent assignment model (top) and independent assignment model (bottom). (a) Example scenario with $P_d = 0.7$ and $\lambda = 10^{-4.5}$. True target detections are red, clutter detections blue. (b) Across 100 instances, we plot the median (solid) and $(0.25, 0.75)$ quantiles (dashed) of L_1 error versus clutter density λ. (c) Close-up track estimates (solid), and one standard deviation error estimates (dashed), for multiple methods applied to a single instance of each dataset.

The measurement residual is calculated based on the nearest-neighbor association as $\nu_{t+1} = y_{t+1}^{\hat{z}_{t+1}} - \hat{y}_{t+1}$. Incorporating the measurement we update the marginal as,

$$\hat{p}_{t+1}(x_{t+1}) = N(x_{t+1} \mid \hat{x}_{t+1|t} + W\nu_{t+1}, \hat{P}_{t+1|t} - W_{t+1}S_{t+1}W_{t+1}^T)$$

where W_{t+1} and S_{t+1} are the typical Kalman gain and innovation covariance, respectively. The smoothed posterior marginal $\hat{p}(x_t \mid Y_1^T)$ is computed as the product of forward and reverse-time filters, using associations as above.

5.1.2 Target Tracking Simulation

We conduct a Monte Carlo simulation for a one-dimensional latent state x_t with random walk dynamics $x_t \sim N(x_{t-1}, \sigma_p^2)$, initialized uniformly in the observation region. Under either assignment model, target detections $y_t^i \sim N(x_t, \sigma_m^2)$ and clutter detections $y_t^i \sim N(0, \sigma_0^2)$. The *clutter density* λ is proportional to the number of false detections.

We evaluate algorithm performance by the L_1 distance from the true posterior

marginals, accurately approximated by finely discretizing the state space and running the forward-backward algorithm for hidden Markov models (HMMs) [169]. This numerical baseline is possible with one-dimensional states, but intractable for higher-dimensional problems where our EP algorithms remain feasible.

We vary the clutter density $\lambda \in \{10^{-5.5}, 10^{-5.0}, 10^{-4.5}, 10^{-4.0}\}$, fixing the probability of detection as $P_d = 0.7$. For every setting of parameters we sample 100 random instances, each with $T = 100$ time points. While convergence is not guaranteed in these models, we achieve adequate convergence for our results by damping the conventional EP message updates [114, 73], with damping parameter $\alpha = 0.5$.

Figure 5.2 shows results for data sampled from the both the dependent and independent assignment models. As measured by L_1 error, EP consistently outperforms baseline methods, and the Gaussian mixture approximation of EPD+ is superior in almost all cases. In general EP seems robust to model mismatch, as EPD+ is effective even for data from the independent assignment model. EPD clearly improves over PDAF.

Figure 5.2(c) shows close-up track estimates for particular instances sampled from each assignment model. KNN consistently underestimates posterior variance, while PDAF overestimates it. State estimates among the EP algorithms are generally comparable or superior to baselines. EPD+ more accurately estimates the posterior variance than other methods.

5.2 Convergent Minimization of Bethe Approximations

Message passing algorithms are convenient for many applications, such as the tracking problems explored in the previous sections. However, neither LBP nor EP are guaranteed to converge. Even in simple continuous models, both methods may improperly estimate invalid or degenerate marginal distributions, such as Gaussians with negative variance. Such degeneracy typically occurs in classes of models for which convergence properties are poor, and there is evidence that these problems are related [111, 38].

Extensive work has gone into developing algorithms which improve on LBP for models with discrete variables, for example by bounding [184, 160] or convexifying [167] the free energy objective. Gradient optimization methods have been applied successfully to binary Ising models [172], but when applied to Gaussian models this approach suffers similar non-convergence and degeneracy issues as LBP. Work on optimization of continuous variational free energies has primarily focused on addressing convergence problems [73]. None of these approaches directly address degeneracy in

the continuous case, and computation may be prohibitively expensive for these direct minimization schemes.

By leveraging gradient projection methods from the extensive literature on constrained nonlinear optimization, we develop an algorithm which ensures that marginal estimates remain valid and normalizable at all iterations. In doing so, we account for important constraints which have been ignored by previous variational derivations of the expectation propagation algorithm [113, 38, 73]. Moreover, by adapting the method of multipliers [22], we guarantee that our inference algorithm converges for most models of practical interest.

5.2.1 Bethe Variational Problems

Recall from Sec. 2.3.2 that fixed points of LBP and EP correspond to stationary points of the constrained Bethe variational problem. We review this formulation for a pairwise MRF,

$$p(x) = \frac{1}{Z} \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t). \tag{5.12}$$

Let $q(x)$ be a *variational distribution* in the exponential family with sufficient statistics $\phi(x)$ and mean parameters $\mu = \mathbb{E}[\phi(x)]$. The variational optimization is over the relaxed constraint set of *locally consistent* marginal distributions $\mathbb{L}(G)$ satisfying expectation constraints associated with each edge of the graph:

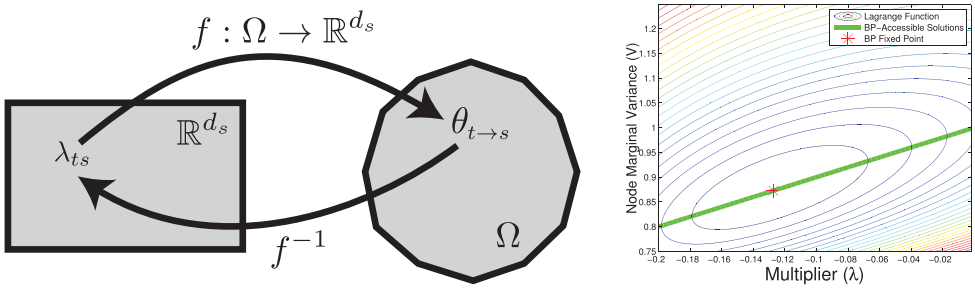
$$C_s(\mu) = 1 - \int q_s(x_s; \mu_s)\, dx_s, \quad C_{ts}(\mu) = \mu_s - \mathbb{E}_{q_{st}}[\phi_s(x_s)].$$

The objective function is the *Bethe free energy*, a tractable approximation to the exact variational free energy that replaces an intractable entropy with that of a tree-structured distribution,

$$\begin{aligned} \mathcal{F}_B(\mu) = &\sum_{(s,t) \in \mathcal{E}} \mathbb{E}_{q_{st}}[\log q_{st}(x_s, x_t) - \log \varphi_{st}(x_s, x_t)] \\ &- \sum_{s \in \mathcal{V}} (n_s - 1)\, \mathbb{E}_{q_s}[\log q_s(x_s) - \log \psi_s(x_s)], \end{aligned} \tag{5.13}$$

where we define the shorthand $\varphi_{st} = \psi_{st}\psi_s\psi_t$ and $n_s = |\Gamma(s)|$ is the number of neighbors to node s. The constrained Bethe variational problem (BVP) is then,

$$\begin{aligned} \underset{\mu}{\text{minimize}} \quad & \mathcal{F}_B(\mu) \\ \text{subject to} \quad & C_{ts}(\mu) = 0, \forall s \in \mathcal{V}, t \in \Gamma(s) \\ & C_s(\mu) = 0, \forall s \in \mathcal{V}, \\ & \{\mu_s : s \in \mathcal{V}\} \cup \{\mu_{st} : (s,t) \in \mathcal{E}\} \in \mathbb{K}. \end{aligned} \tag{5.14}$$



Lagrange / Message Mapping Gaussian Lagrange Function

Figure 5.3: **Correspondence to Bethe free energy.** *Left:* For stationary points of the Lagrangian (5.15) a bijection f exists between multipliers λ and canonical parameters θ of the variational distribution. *Right:* Level sets of the Lagrangian for a Gaussian MRF. Contours show the function in terms of the lagrange multiplier and marginal variance, the marginalization constraint (green line) and unique BP fixed point (red star) are also shown.

The constraint set $\mathbb{K} = \bigcup_s \mathbb{K}_s \bigcup_{st} \mathbb{K}_{st}$ defines the set of valid mean parameters for each node μ_s and edge μ_{st} marginal. The definition of $\mathbb{K}$ depends on the variational distribution q, for example if q is Gaussian then $\mathbb{K}$ is the positive semidefinite cone.

Correspondence to Message Passing

We can optimize the BVP (5.14) by relaxing the normalization and local consistency constraints with Lagrange multipliers. Constrained minima are characterized by stationary points of the Lagrangian,

$$\mathcal{L}(\mu, \lambda) = \mathcal{F}_B(\mu) + \sum_s \lambda_s C_s(\mu) + \sum_s \sum_{t \in N(s)} \lambda_{ts} C_{ts}(\mu). \tag{5.15}$$

Equivalence between LBP fixed points and stationary points of the Lagrangian for the discrete case have been discussed extensively [183, 169]. Similar correspondence has been shown more generally for EP fixed points [169, 74]. Since our focus is on the continuous case we briefly review the correspondence between Gaussian LBP fixed points and the Gaussian Bethe free energy. For simplicity we focus on zero-mean $p(x) = N(x \mid 0, J^{-1})$, where diagonal precision entries $J_{ss} = A_s$ and

$$\psi_s(x_s) = \exp\left\{-\frac{1}{2} x_s^2 A_s\right\}, \qquad \psi_{st}(x_s, x_t) = \exp\left\{-\frac{1}{2} \begin{pmatrix} x_s & x_t \end{pmatrix} \begin{pmatrix} 0 & J_{st} \\ J_{st}^T & 0 \end{pmatrix} \begin{pmatrix} x_s \\ x_t \end{pmatrix}\right\}.$$

Let $q(x_s) = N(x_s \mid 0, V_s)$, $q(x_s, x_t) = N(\left(\begin{smallmatrix} x_s \\ x_t \end{smallmatrix}\right) \mid 0, \Sigma_{st})$, $\Sigma_{st} = \left(\begin{smallmatrix} V_{ts} & P_{st} \\ P_{ts} & V_{st} \end{smallmatrix}\right)$, and $\widetilde{B}_{st} = \left(\begin{smallmatrix} A_s & J_{st} \\ J_{st} & A_t \end{smallmatrix}\right)$. The Gaussian Bethe free energy then equals:

$$\mathcal{F}_{GB}(V, \Sigma) = \frac{1}{2} \sum_{(s,t) \in \mathcal{E}} \left(\operatorname{tr}(\Sigma_{st} \widetilde{B}_{st}) - \log |\Sigma_{st}|\right) - \sum_{s \in \mathcal{V}} \left(\frac{n_s - 1}{2}\right) (V_s A_s - \log V_s). \tag{5.16}$$

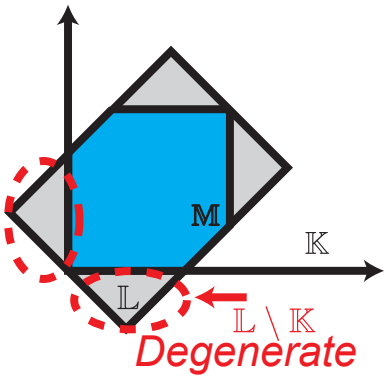


Figure 5.4: **Degenerate mean parameters.** The constraints $\mathbb{M}$ defining the marginal polytope are a subset of the locally consistent marginals defined by $\mathbb{L}$. For continuous MRFs $\mathbb{L}$ may include marginals outside the set of valid mean parameters $\mathbb{K}$, e.g. Gaussians with negative definite covariance.

The locally consistent marginal polytope $\mathbb{L}(G)$ consists of the constraints $C_{ts}(V) = V_s - V_{ts}$ for all nodes $s \in \mathcal{V}$ and edges $(s,t) \in \mathcal{E}$. The Lagrangian is given by,

$$\mathcal{L}(V, \Sigma, \lambda) = \mathcal{F}_{GB}(V, \Sigma) + \sum_s \sum_{t \in N(s)} \lambda_{ts} \left[V_s - V_{ts} \right]. \tag{5.17}$$

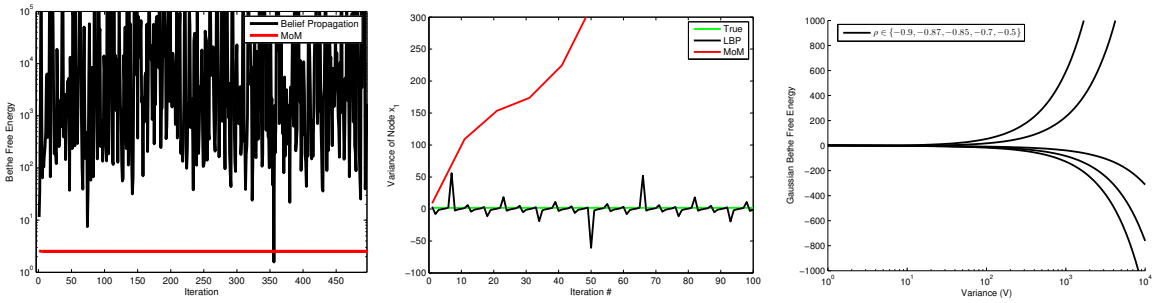
Taking the derivative with respect to the node marginal variance $\frac{\partial \mathcal{L}}{\partial V_s} = 0$ yields the stationary point $V_s^{-1} = A_s + \frac{1}{n_s - 1} \sum_{t \in N(s)} \lambda_{ts}$. For a Gaussian LBP algorithm with messages parameterized as $m_{t \to s}(x_s) = \exp\left\{-\frac{1}{2} x_s^2 \Lambda_{t \to s}\right\}$, fixed points of the node marginal precision are given by

$$\Lambda_s = A_s + \sum_{t \in N(s)} \Lambda_{t \to s}$$

Let $\lambda_{ts} = -\frac{1}{2} \sum_{a \in N(s) \setminus t} \Lambda_{a \to s}$. Substituting back into the stationary point conditions yields $V_s^{-1} \Rightarrow \Lambda_s$. A similar construction holds for the pairwise marginals. Inverting the correspondence between multipliers and message parameters yields the converse $V_s^{-1} \Leftarrow \Lambda_s$ (c.f. [74]).

Message Passing Non-Convergence and Degeneracy

While local message passing algorithms are convenient for many applications, their convergence is not guaranteed in general. Loopy BP, for example, often fails to converge for networks with tight loops [183] such as the 3×3 lattice of Figure 5.5(a). For non-Gaussian models with continuous variables, convergence of the EP algorithm can be even more problematic [73].



(a) (b) (c)

Figure 5.5: (a) Bethe free energy versus iteration for 3x3 toroidal binary MRF. (b) Node marginal variance estimates per iteration for a symmetric, single-cycle Gaussian MRF with three nodes (plot is of x_1, other nodes are similar). (c) For the model from (b), the Gaussian Bethe free energy is unbounded on the constraint set.

For continuous models message updates may yield degenerate, unnormalizable marginal distributions which do not correspond to stationary points of the Lagrangian. For example, for Gaussian MRFs the Bethe free energy $\mathcal{F}_B(\cdot)$ is derived from expectations with respect to variational distributions over nodes $q_s(x_s; \mu_s)$ and edges $q_{st}(x_s, x_t; \mu_{st})$. If a set of hypothesized marginals are not normalizable (positive variance), the Gaussian Bethe free energy $\mathcal{F}_{GB}(\cdot)$ is invalid and undefined.

Degenerate marginals arise because the constraint $\mathbb{K}$ over valid mean parameters is not represented in the Lagrangian (5.15); this issue is mentioned briefly in [169] but is not dealt with computationally. Figure 5.5(b) demonstrates this issue for a simple, three-node Gaussian MRF. Here LBP produces marginal variances which oscillate between impossibly large positive, and non-sensical negative, values. Such degeneracies are arguably more problematic for EP since its *moment matching* steps require expected values with respect to an *augmented distribution* [114], which may involve an unbounded integral.

Unboundedness of the Gaussian Bethe Free Energy

Conditions under which the simple LBP and EP updates are guaranteed to be accurate are of great practical interest. For Gaussian MRFs, the class of *pairwise normalizable* models are sufficient to guarantee LBP stability and convergence [111]. For non-pairwise normalizable models the Gaussian Bethe free energy is unbounded below [38] on the set of local consistency constraints $\mathbb{L}(G)$.

We offer a small example consisting of a non-pairwise normalizable symmetric single cycle with 3 nodes. Diagonal precision elements are $J_{ss} = 1.0$, and off-diagonal

elements $J_{st} = 0.6$. We embed marginalization constraints into a symmetric parameterization $V_s = V$ and $\Sigma_{st} = \left(\begin{smallmatrix} V & \rho V \\ \rho V & V \end{smallmatrix}\right)$. Feasible solutions within the constraint set are characterized by $V > 0$ and $-1 < \rho < 1$. Substituting this parameterization into the Gaussian free energy (5.16), and performing some simple algebra, yields

$$\mathcal{F}_{GB}(V, \rho) = -\frac{3}{2}\log V + \frac{3}{2}V(1 + 1.2\rho) - \frac{3}{2}\log(1 - \rho^2).$$

For $\rho < -\frac{1}{1.2}$ the free energy is unbounded below at rate $\mathcal{O}(-V)$. Figure 5.5(c) illustrates the Bethe free energy for this model as a function of V, and for several values of ρ.

More generally, it has been shown that Gaussian EP messages are always normalizable (positive variance) for models with log-concave potentials [142]. It has been conjectured, but not proven, that EP is also guaranteed to converge for such models [136]. For Gaussian MRFs, we note that the family of log-concave models coincides with the pairwise normalizability condition. Our work seeks to improve inference for non-log-concave models with bounded Bethe free energies.

5.2.2 Method of Multipliers (MoM) Optimization

Given our complete constrained formulation of the Bethe variational problem, we avoid convergence and degeneracy problems via direct minimization using the *method of multipliers* (MoM) [22]. In general terms, given some convex feasible region $\mathbb{K}$, consider the equality constrained problem

$$\underset{x \in \mathbb{K}}{\text{minimize}}\ f(x) \quad \text{subject to} \quad h(x) = 0$$

With penalty parameter $c > 0$, we form the *augmented Lagrangian* function,

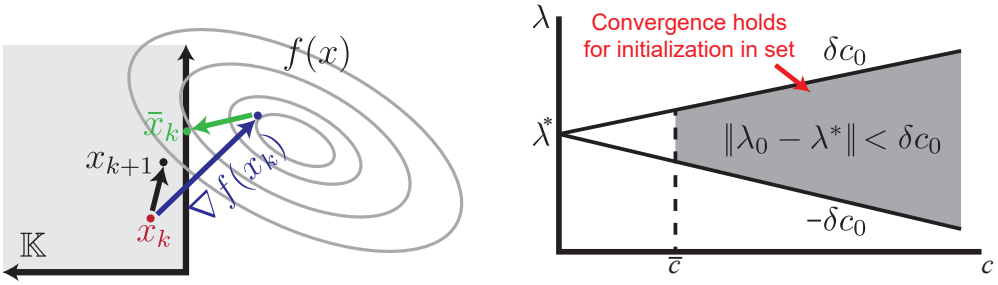
$$\mathcal{L}_c(x, \lambda) = f(x) + \lambda^T h(x) + \frac{1}{2}c||h(x)||^2 \tag{5.18}$$

Given a multiplier vector λ_k and penalty parameter c_k we update the primal and dual variables as,

$$x_k = \arg\min_{x \in \mathbb{K}} \mathcal{L}_{c_k}(x, \lambda_k), \quad \lambda_{k+1} = \lambda_k + c_k h(x_k). \tag{5.19}$$

The penalty multiplier can be updated as $c_{k+1} \geq c_k$ according to some fixed update schedule, or based on the results of the optimization step. An update rule that we find useful [22] is to increase the penalty parameter by $\beta > 1$ if the constraint violation is not improved by a factor $0 < \gamma < 1$ over the previous iteration,

$$c_{k+1} = \begin{cases} \beta c_k & \text{if } \|h(x_k)\| > \gamma\|h(x_{k-1})\|, \\ c_k & \text{if } \|h(x_k)\| \leq \gamma\|h(x_{k-1})\|. \end{cases} \tag{5.20}$$



Gradient Projection MoM Convergence

Figure 5.6: **Bethe optimization.** *Left:* Gradient projection enforces feasible marginal mean parameters. Each iteration the gradient $\nabla f(x_k)$ is calculated and Euclidean projection enforces the closest feasible solution $\bar{x}_k$. Iterates x_{k+1} are generated in the direction of the vector $(\bar{x}_k - x_k)$. *Right:* Method of multiplier optimization is guaranteed to converge for a given initialization (c_0, λ_0) provided the suboptimality of the initial multiplier λ_0 is upper bounded by the initial penalty c_0 by a linear coefficient $\delta > 0$: $\|\lambda_0 - \lambda^*\| < \delta c_0$.

Gradient Projection

The augmented Lagrangian $\mathcal{L}_c(x, \lambda)$ is a partial one, where feasibility of mean parameters ($x \in \mathbb{K}$) is enforced explicitly by projection. A simple *gradient projection* method [22] defines a sequence

$$x_{k+1} = x_k + \alpha_k(\bar{x}_k - x_k), \qquad \bar{x}_k = [x_k - s_k \nabla f(x_k)]^+ . \tag{5.21}$$

The notation $[\cdot]^+$ denotes a projection onto the constraint set $\mathbb{K}$. After taking a step $s_k > 0$ in the direction of the negative gradient, we project the result onto the constraint set to obtain a feasible direction $\bar{x}_k$. We then compute x_{k+1} by taking a step $\alpha_k \in (0, 1]$ in the direction of $(\bar{x}_k - x_k)$. If $x_k - s_k \nabla f(x_k)$ is feasible, gradient projection reduces to unconstrained steepest descent.

There are multiple such projection steps in each *inner-loop* iteration of MoM (e.g. each x_k update). For our experiments we use a projected quasi-Newton method [141] and step-sizes α_k and s_k are chosen using an Armijo rule [22, Prop. 2.3.1].

Convergence Properties

Convergence and rate of convergence results have been proven [21, Proposition 2.4] for the Method of Multipliers with a quadratic penalty and multiplier iteration $\lambda_{k+1} = \lambda_k + c_k h(x_k)$. The main regularity assumptions are that the sequence $\{\lambda_k\}$ is bounded, and there is a local minimum for which a Lagrange multiplier pair (x^*, λ^*) exists satisfying second-order sufficiency conditions, so that $\nabla_x \mathcal{L}_0(x^*, \lambda^*) = 0$ and $z^T \nabla^2_{xx} \mathcal{L}_0(x^*, \lambda^*) z > 0$ for all $z \neq 0$. It then follows that there exists some $\bar{c}$ such that for all $c \geq \bar{c}$, the augmented Lagrangian also contains a strict local minimum $z^T \nabla^2_{xx} \mathcal{L}_c(x^*, \lambda^*) z > 0$.

For convergence, the initialization of the Lagrange multiplier λ_0 and penalty parameter c_0 must be such that $\|\lambda_0 - \lambda^*\| < \delta c_0$ for some $\delta > 0$ and $c \geq \bar{c}$ which depend on the objective and constraints. In practice, a poor initialization of the multiplier λ_0 can often be offset by a sufficiently high c_0. A final technical note is that convergence proofs assume the sequence of unconstrained optimizations which yield x_k stays in the neighborhood of x^* after some k. This does not hold in general, but can be encouraged by warm-starting the unconstrained optimization with the previous x_{k-1}.

To invoke existing convergence results we must show that a local minimum x^* exists for each of the free energies we consider; a sufficient condition is then that the Bethe free energy is bounded from below. This property has been previously established for general discrete MRFs [72], for pairwise normalizable Gaussian MRFs [38], and for the clutter model [114]. For non-pairwise normalizable Gaussian MRFs, the example of Section 5.2.1 shows that the Bethe variational objective is unbounded below, and further may not contain any local optima. While the method of multipliers does not converge in this situation, its non-convergence is due to fundamental flaws in the Bethe approximation.

5.2.3 MoM Algorithms for Probabilistic Inference

We derive MoM algorithms which minimize the Bethe free energy for three different families of graphical models. For each model we define the form of the joint distribution, Bethe free energy $\mathcal{F}_{\mathrm{B}}$, local consistency constraints, augmented Lagrangian, and the gradient projection step. Gradients, which can be notationally cumbersome, are given in Appendix A.1.

Gaussian Markov Random Fields

We have already introduced the Lagrangian (5.17) for the Gaussian MRF. The Gaussian Bethe free energy (5.16) is always unbounded below off of the constraint set in node marginal variances V_s. We correct this by adding an additional fixed penalty in the augmented Lagrangian,

$$\begin{aligned}\mathcal{L}_c(V, \Sigma, \lambda) = \mathcal{F}_{GB}(V) &+ \sum_s \sum_{t \in N(s)} \lambda_{ts} \left[V_s - V_{ts}\right] \\ &+ \frac{\kappa}{2} \sum_s \sum_{t \in N(s)} \left[\log V_s - \log V_{ts}\right]^2 + \frac{c}{2} \sum_s \sum_{t \in N(s)} \left[V_s - V_{ts}\right]^2 .\end{aligned}$$

We keep $\kappa \geq 1$ fixed so that existing convergence theory remains applicable. The set of realizable mean parameters $\mathbb{K}$ is the set of symmetric positive semidefinite matrices V_s, Σ_{st}. We therefore must solve a series of constrained optimizations of the

form, $\min_{V,\Sigma} \mathcal{L}_{c_k}(V, \Sigma, \lambda_k)$, subject to $V_s \geq 0, \Sigma_{st} \succeq 0$. The gradient projection step is easily expressed in terms of correlation coefficients ρ_{st},

$$\Sigma_{st} = \begin{bmatrix} V_{st} & \rho_{st}\sqrt{V_{st}V_{ts}} \\ \rho_{st}\sqrt{V_{st}V_{ts}} & V_{ts} \end{bmatrix}.$$

Then, $\Sigma_{st} \succeq 0$ if and only if $V_{st} \geq 0$, $V_{ts} \geq 0$, and $-1 \leq \rho_{st} \leq 1$. The projection step is then,

$$V_{st} = \max(0, V_{st}), \quad V_{ts} = \max(0, V_{ts}), \quad \rho_{st} = \max(-1, \min(1, \rho_{st})).$$

The full MoM algorithm follows from gradient derivations in Appendix A.1.

Recall that in Section 5.2.1, we showed that the Gaussian Bethe free energy is unbounded on the constraint set for non-pairwise normalizable models. We run MoM on the symmetric three-node cycle from this discussion and find that MoM, correctly, identifies an unbounded direction, and Figure 5.5(b) shows that the node marginal variances indeed diverge to infinity.

Discrete Markov Random Fields

Consider a pairwise MRF where all variables $x_s \in \mathcal{X}_s = \{1, \ldots, K_s\}$ are discrete. The variational marginal distributions are then $q_s(x_s; \tau) = \prod_{k=1}^{K_s} \tau(x_s)^{\mathbb{I}(x_s,k)}$, and have mean parameters $\tau \in \mathbb{R}^{K_s}$. Let $\tau(x_s)$ denote element x_s of vector τ. Pairwise marginals have mean parameters $\tau_{st} \in \mathbb{R}^{K_s \times K_t}$ similarly indexed as $\tau_{st}(x_s, x_t)$. The discrete Bethe free energy is then

$$\begin{aligned} \mathcal{F}_B(\tau; \psi) = &\sum_{(s,t)\in\mathcal{E}} \sum_{x_s} \sum_{x_t} \tau_{st}(x_s, x_t)[\log \tau_{st}(x_s, x_t) - \log \psi_{st}(x_s, x_t)] \\ &- \sum_{s\in\mathcal{V}} \sum_{x_s} (n_s - 1)\tau_s(x_s)[\log \tau_s(x_s) - \log \psi_s(x_s)]. \end{aligned}$$

For this discrete model, our expectation constraints reduce to the following normalization and marginalization constraints:

$$C_s(\tau) = 1 - \sum_{x_s} \tau_s(x_s), \quad C_{ts}(x_s; \tau) = \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t).$$

The augmented Lagrangian is then,

$$\begin{aligned} \mathcal{L}_c(\tau, \lambda, \xi; \psi) = \mathcal{F}_B(\tau; \psi) + \sum_{(s,t)\in\mathcal{E}} \left[\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s; \tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; \tau) \right] \\ + \sum_{s\in\mathcal{V}} \xi_{ss} C_s(\tau) + \frac{c}{2} \sum_{s\in\mathcal{V}} C_s(\tau)^2 + \frac{c}{2} \sum_{(s,t)\in\mathcal{E}} \left[\sum_{x_s} C_{ts}(x_s; \tau)^2 + \sum_{x_t} C_{st}(x_t; \tau)^2 \right]. \end{aligned}$$

Mean parameters must be non-negative to be valid, so $\mathbb{K} = \{\tau_s, \tau_{st} : \tau_s \geq 0, \tau_{st} \geq 0\}$. This constraint is enforced by a bound projection $\tau_s(x_s) = \max(0, \tau_s(x_s))$, and similarly for the pairwise marginals. While these constraints are never active in BP fixed point iterations, they must be enforced in gradient optimization. With these pieces and the gradient computations presented in Appendix A.1, implementation of MoM optimization for the discrete MRF is straightforward.

Discrete Mixtures of Gaussian Potentials

We are particularly interested in tractable inference in hybrid models with discrete and conditionally Gaussian random variables. A simple example of such a model is the *clutter problem* [114], whose joint distribution models N conditionally independent Gaussian observations $\{y_i\}_{i=1}^N$. These observations may either be centered on a target scalar $x \in \mathbb{R}$ ($z_i = 1$) or drawn from a background clutter distribution ($z_i = 0$). If target observations occur with frequency β_0, we then have

$$x \sim N(\mu_0, P_0), \quad z_i \sim \text{Ber}(\beta_0), \quad y_i \mid x, z_i \sim N(0, \sigma_0^2)^{(1-z_i)} N(x, \sigma_1^2)^{z_i}$$

The corresponding variational posterior distributions are,

$$q_0(x) = N(m_0, V_0), \quad q_i(x, z_i) = \left((1-\beta_i) N(x \mid m_{i0}, V_{i0})\right)^{(1-z_i)} \left(\beta_i N(x \mid m_{i1}, V_{i1})\right)^{z_i}.$$

Assuming normalizable marginals with $V_0 \geq 0$, $V_{i0} \geq 0$, $V_{i1} \geq 0$, as always ensured by our multiplier method, we define the Bethe free energy $\mathcal{F}_{CGB}(m, V, \beta)$ in terms of the mean parameters in Appendix A.1. Expectation constraints are given by,

$$C_i^{\text{mean}} = \mathbb{E}_0[x] - \mathbb{E}_i[x], \quad C_i^{\text{var}} = \text{Var}_0[x] - \text{Var}_i[x],$$

where $\mathbb{E}_i[\cdot]$ and $\text{Var}_i[\cdot]$ denote the mean and variance of the Gaussian mixture $q_i(x, z_i)$. Combining the free energy, constraints, and additional positive semidefinite constraints on the marginal variances we have the BVP for the clutter model,

$$\begin{aligned} \underset{m,V,\beta}{\text{minimize}} \quad & \mathcal{F}_{CGB}(m, V, \beta; \psi) \\ \text{subject to} \quad & C_i^{\text{mean}} = 0, \; C_i^{\text{var}} = 0, \; \text{for all } i = 1, 2, \ldots, N \\ & V_0 \geq 0, V_{i0} \geq 0, V_{i1} \geq 0 \end{aligned} \tag{5.22}$$

Derivation of the free energy and augmented Lagrangian is somewhat lengthy, and so is deferred to Appendix A.1. Projection of the variances onto the constraint set is a simple thresholding operation.

5.2.4 Experimental Results

5.2.5 Discrete Markov Random Fields

We consider binary Ising models, with variables arranged in NxN lattices with toroidal boundary conditions. Potentials are parameterized as in [182], so that

$$\psi_s = \begin{bmatrix} \exp(h_s) \\ \exp(-h_s) \end{bmatrix}, \quad \psi_{st} = \begin{bmatrix} \exp(J_{st}) & \exp(-J_{st}) \\ \exp(-J_{st}) & \exp(J_{st}) \end{bmatrix}.$$

We sample 500 instances at random from a 10x10 toroidal lattice with each $J_{st} \sim N(0,1)$ and $h_s \sim N(0, 0.01)$. LBP is run for a maximum of 1000 iterations, and MoM is initialized with a single iteration of LBP. We report average L_1 error of the approximate marginals as compared to the true marginals computed with the junction tree algorithm [117]. Marginal errors are reported in Figure 5.7(a,top), and there is a clear improvement over LBP in the majority of cases.

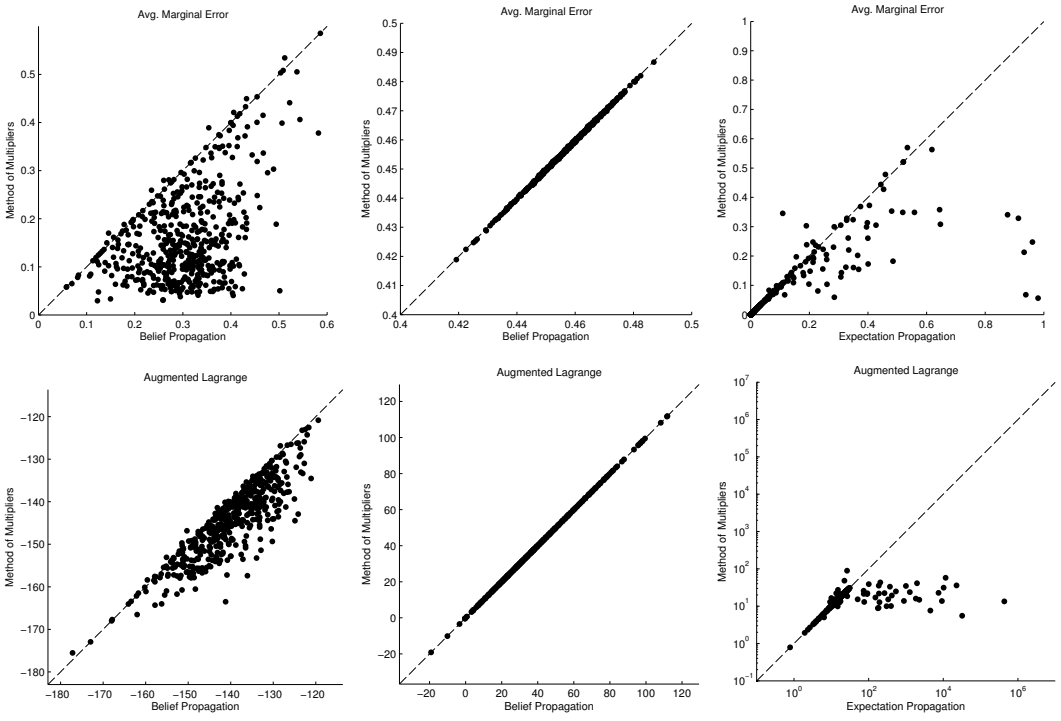
Direct evaluation of the Bethe free energy does not take into account constraint violations for non-convergent LBP runs. The augmented Lagrangian penalizes constraint violation, but requires a penalty parameter which LBP does not provide. For an objective comparison, we construct a penalized Bethe free energy by evaluating the augmented Lagrangian with fixed penalty $c = 1$ and multipliers $\lambda = 0$. We evaluate this objective at the final iteration of both algorithms. As we see in Figure 5.7(a,bottom), MoM finds a lower free energy for most trials.

Our implementations of LBP and MoM are in Matlab, and emphasize correctness over efficiency. Nevertheless, computation time for LBP exceeds that of MoM. Wall clock time is measured in seconds across various trials, and the percentiles for LBP are 25%: 1040.46, 50%: 1042.57, and 75%: 1044.85. For MoM they are 25%: 290.25, 50%: 381.62, and 75%: 454.52.

Gaussian Markov Random Fields

For the Gaussian case we again sample 500 random instances from a 10x10 lattice with toroidal boundary conditions. We randomly sample only pairwise normalizable instances and initialization is provided with a single iteration of Gaussian LBP. We find that MoM is generally insensitive to initialization in this model. True marginals are computed by explicitly inverting the model precision matrix and average symmetric L_1 error with respect to truth is reported in Figure 5.7(b,top).

For pairwise normalizable models, Gaussian LBP is guaranteed to converge to the unique fixed point of the Bethe free energy, so it is reassuring that MoM optimization matches LBP performance. The value of the augmented Lagrangian at the final iteration is shown in Figure 5.7(b,bottom) and again shows that MoM matches Gaussian



(a) (b) (c)

Figure 5.7: Performance of MoM and LBP on randomly generated (a) discrete 10×10 toroidal Ising MRFs, (b) 10×10 toroidal Gaussian MRFs, and (c) clutter models with $N = 30$ observations. Each point corresponds to a single model instance. *Top:* L_1 error between estimated and true marginal distributions, averaged over all nodes. *Bottom:* Penalized Bethe free energy constructed by setting $\lambda = 0$, $c = 1$ in the augmented Lagrangian.

LBP on pairwise normalizable models. Computation time for MoM is slightly faster with median wall clock time of 58.76 seconds as compared to 103.17 seconds for LBP. The 25% and 75% percentiles are 37.81 and 92.10 seconds for MoM compared to 88.40 and 125.59 seconds for LBP.

Discrete Mixtures of Gaussian Potentials

To test the benefits of avoiding degenerate marginals, we consider the clutter model of Sec. 5.2.3 with $\mu_0 = 0$, $P_0 = 100$ and $\beta_0 = 0.25$. The variance of the clutter distribution is $\sigma_0^2 = 10$, and of the target distribution $\sigma_1^2 = 1$. We sample $N = 30$ observations for each trial instance.

A good initialization of the multipliers is critical to performance of MoM. We generate 10 initializations by running 5 iterations of EP, each with a different random message update schedule, compute the corresponding Lagrange multipliers for

each, and use the one with the lowest value of the augmented Lagrangian. Similarly, we measure EP's performance by the best performing of 10 longer runs. Both methods are run for a maximum of 1000 iterations, and true marginals are computed numerically by finely discretizing the scalar target x.

We sample 500 random instances and report average L_1 error with respect to true marginals in Figure 5.7(c,top). We see a significant improvement in the majority of runs. Similarly, the augmented Lagrangian comparison is shown in Figure 5.7(c,bottom) and MoM often finds a better penalized free energy. While MoM and EP can both suffer from local optima, MoM avoids non-convergence and the output of invalid (negative variance) marginal distributions. Median wall clock time for EP is 0.59 seconds, and 9.80 seconds for MoM. The 25% and 75% percentiles are 0.42 and 0.84 seconds for EP and 0.51 and 49.19 seconds for MoM.

5.3 Discussion

We began this chapter by developing a family of target tracking algorithms which are significantly more accurate than baseline methods (Sec. 5.1). Contrary to standard target tracking approaches, which make explicit choices for the distribution of false detections, our method allows flexibility in underlying observation model. This flexibility enables a tradeoff between accuracy and computation with a common approach to inference based on expectation propagation.

Although the underlying EP inference remains similar across models, we observe different stability and convergence behavior when varying model complexity. Motivated by these degeneracies we investigate an approach for directly minimizing the Bethe variational problem underlying EP message passing (Sec. 5.2). Our approach is unique in that we do not relax the constraint on normalizability of the marginals, rather we explicitly enforce it at all points in the optimization. This method directly avoids the creation of degenerate distributions — for example with negative variance — which frequently occur in more greedy approaches for minimizing the Bethe free energy. Moreover, we obtain convergence guarantees under broadly applicable assumptions, thereby avoiding one practical limitation of EP. We further note that our optimization of the Bethe variational problem is not specific to the models we have chosen to investigate, but rather can be extended to arbitrary MRFs.

Chapter 6

Contributions and Suggestions

The applications we explore in this thesis span a broad range from signal processing and computer vision to computational biology. The statistical inference algorithms we develop, though, generalize easily to other problem domains. In this chapter we review the main contributions of this thesis and conclude with a discussion of further research directions.

6.1 Discussion of Contributions

A ubiquitous adoption of probabilistic graphical models by the machine learning community has led to models of ever-increasing complexity. These models capture the statistics of complicated processes, but they pose a challenge for statistical inference algorithms. In this thesis we develop inference algorithms that can flexibly adapt to models of arbitrary complexity. We develop techniques for both marginal posterior and MAP, and explore them in a variety of contexts.

Much of our contribution is outlined in Chapter 3, where we develop particle-based MAP inference for continuous MRFs. Our primary contribution is a method which maintains solutions at multiple local modes of the distribution, thereby remaining robust to initialization and model mismatch. We formulate the Diverse Particle Max-Product (D-PMP) algorithm along with variations of existing particle-based approximations to max-product BP.

Articulated physical models are convenient for demonstrating the need for solution diversity. The local nature of the model definition gives rise to global ambiguities. In models of human pose estimation, for example, image likelihoods can be noisy and uninformative. Solution hypotheses can easily be visualized and qualitatively assessed for consistency. Interpreting these ambiguities can be difficult for models of other types of phenomena, but as we highlight in the optical flow experiments of Chapter 3

they are still very important. While estimates of optical flow are unambiguous in homogeneous regions, they can pose challenges around object boundaries and in the presence of significant occlusion. In this setting D-PMP is competitive against a highly engineered method which has been tuned to the particular dataset we consider.

A different notion of diversity arises in the context of protein structure prediction, which we explore in Chapter 4, whereby the underlying physical system does not exist in thermal equilibrium. As a result it is widely accepted that proteins often assume multiple stable conformations, often resulting in different functions. Most approaches capture multiple conformations indirectly, by running parallel inference chains from random initializations. Moreover, the high dimensionality of the latent space necessitates discretization, for example via rotamer libraries used in side chain prediction. By contrast, D-PMP both minimizes the continuous energy model and preserves multiple stable configurations in the particle set. Using this approach D-PMP yields results that are more accurate than highly engineered methods based on simulated annealing Monte Carlo.

The second half of this thesis develops variational methods for marginal inference. Chapter 5 begins with an exploration of variational approximations in time-series models. Through the application of target tracking, in the presence of false detections, we investigate efficacy of EP inference for various observation models. Using this approach we are able to generalize several existing tracking methods in the literature, and develop novel methods.

In our target tracking investigation we observe classic degeneracies whereby EP produces unnormalizable marginal approximations, or does not converge at all. Motivated by these failures we investigate the Bethe variational optimization underlying EP and BP. Using techniques from nonlinear optimization we develop an approach which directly minimizes the Bethe free energy. In continuous models this approach directly enforces normalizability constraints, and so is guaranteed to produce sensible marginal approximations. The method is also guaranteed to converge, under mild assumptions, and in discrete MRFs we observe that it often produces more accurate marginal estimates than loopy BP when message passing convergence cannot be achieved.

6.2 Suggestions for Future Research

The following subsections present a concise list of future directions which build on the work presented in this thesis. We briefly discuss how our methods can be applied to learn hyperparameters in structured models, to exploit solution diversity in continuous models, and to improve the state-of-the-art in estimating protein folding.

6.2.1 Exploiting Solution Diversity

Reranking plausible hypotheses is a popular technique for boosting model accuracy by incorporating high-order potentials, which cannot be included in a tractable model. Such external reasoning is responsible for state-of-the-art results in natural language parsing [31, 34], machine translation [66], image segmentation and human pose estimation [131, 175]. In protein structure prediction too, reranking solutions from competing algorithms accounts for the leading results in biennial competitions [118].

Exact methods for generating the *M-best MAP* solutions scale exponentially with tree width [102, 121, 144, 138]. However, approximations have been proposed based on successive calls to MP inference [180], LP relaxations [59] or sampling [15, 164].

The diverse particle selection methods proposed in Chapter 3 preserve diversity among particles, but the M-best MAP solutions often remain similar. The *diverse M-best MAP* instead yields a set of high probability solutions with enforced dissimilarity [17]. At present current investigations of diverse M-best MAP solutions have been limited to discrete MRFs. An interesting line of work exists to explore the connections between diverse particle sets on continuous MRFs and diverse solution sets on the particle discretization.

6.2.2 Structured Learning of Continuous MRFs

Increasing model complexity results in large numbers of parameters that must be learned. Structured support vector machines (S-SVMs) offer one approach for training discrete MRFs by extending the *max-margin* property of classical SVMs to multivariate discrete outputs. Additionally, S-SVM is formulated as a convex optimization with a similar structure to classical SVMs.

The S-SVM framework, however, does not extend straightforwardly to MRFs over continuous-valued random variables, such as the ones we consider in this thesis. Recent work in Bayesian optimization takes a different approach by performing Gaussian process regression on the performance function of the algorithm [152]. This approach lacks the generalization properties that max-margin learning affords.

Structured SVM optimization based on subgradient [98] or cutting plane [86] approaches generally invoke an *oracle* which solves a modified MAP problem. Repeated D-PMP inference approximates continuous S-SVM learning via a succession of discretization of the S-SVM sub-problems. In this way S-SVM can be elegantly extended to the continuous domain while maintaining generalization attributes similar to those in the discrete setting.

6.2.3 Particle Representations for Protein Folding

The success of D-PMP inference for protein side chain prediction (see Chapter 4) suggests further gains can be realized by extending particle max-product to full protein folding. Estimating the protein backbone is a significantly more complex problem, yet state-of-the-art protein folding algorithms rely on simulated annealing optimization similar to those in the simpler side chain prediction task.

Estimation of the protein backbone introduces several complications not present in side chain prediction. Firstly, a particle representation must be chosen which minimizes the number of degrees of freedom but that is sufficiently expressive. Traditional dihedral angle representations require a state-space that spans multiple neighboring residues, thereby constraining exploration. Secondly, the fixed backbone in side chain prediction allows us to specify MRF edges based on proximity between residues, but these relative distances continually evolve as backbone estimates are refined. An analogous representation for protein folding would require a dynamic MRF topology that is updated with each stage of inference.

Appendix A

Derivations and Proofs

A.1 Gradient Calculations for Bethe Minimization

Section 5.2.3 provides an overview of augmented Lagrangian methods for minimizing Bethe free energies. The main text formulates the objective, Lagrangian, and projection operators for MRFs with discrete, Gaussian, and Gaussian mixture potentials. This appendix details gradients of the augmented Lagrangian in these models, which are necessary for implementation.

A.1.1 Discrete Markov Random Fields

Recall the Bethe free energy for a discrete pairwise MRF is given by,

$$\begin{aligned}\mathcal{F}_B(\tau;\psi) = &\sum_{(s,t)\in\mathcal{E}} \sum_{x_s}\sum_{x_t} \tau_{st}(x_s,x_t)[\log \tau_{st}(x_s,x_t) - \log \psi_{st}(x_s,x_t)] \\ &- \sum_{s\in\mathcal{V}} \sum_{x_s} (n_s-1)\tau_s(x_s)[\log \tau_s(x_s) - \log \psi_s(x_s)],\end{aligned}$$

where τ_s and τ_{st} are the marginal mean parameters for node and edge marginals, respectively, and $n_s = |\Gamma(s)|$ the number of neighbors for node s. Normalization and marginalization constraints are given by,

$$C_s(\tau) = 1 - \sum_{x_s} \tau_s(x_s), \quad C_{ts}(x_s;\tau) = \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s,x_t).$$

Let λ_s be the Lagrange multiplier penalizing the normalization constraint and λ_{ts} the multiplier for marginalization. Combining the Bethe free energy and constraints we

specify the augmented Lagrangian with quadratic penalty parameter $c \geq 1$:

$$\mathcal{L}_c(\tau, \lambda; \psi) = \mathcal{F}_B(\tau; \psi) + \sum_{(s,t)\in\mathcal{E}} \left[\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s; \tau) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t; \tau) \right]$$
$$+ \sum_{s\in\mathcal{V}} \lambda_s C_s(\tau) + \frac{c}{2} \sum_{s\in\mathcal{V}} C_s(\tau)^2 + \frac{c}{2} \sum_{(s,t)\in\mathcal{E}} \left[\sum_{x_s} C_{ts}(x_s; \tau)^2 + \sum_{x_t} C_{st}(x_t; \tau)^2 \right].$$

The gradients can be expressed more compactly by first defining the discrete BP fixed points given by [183],

$$\tau_s^{BP}(x_s; \lambda) = \varphi_s(x_s) \exp\left\{ \frac{1}{n_s - 1} \sum_{t\in N(s)} \lambda_{ts}(x_s) \right\} \tag{A.1}$$
$$\tau_{st}^{BP}(x_s, x_t; \lambda) = \phi_{st}(x_s, x_t) \exp\Big\{ \lambda_{ts}(x_s) + \lambda_{st}(x_t) \Big\}. \tag{A.2}$$

The gradients then take the intuitive form,

$$\frac{\partial \mathcal{L}_c}{\partial \tau_s(x_s)} = (n_s - 1)\Big[\log \tau_s^{BP}(x_s) - \log \tau_s(x_s) - 1 \Big] - \lambda_s + c\left[C_{ts}(x_s; \tau) - C_s(\tau) \right]$$
$$\frac{\partial \mathcal{L}_c}{\partial \tau_{st}(x_s)} = \log \tau_{st}(x_s, x_t) + 1 - \log \tau_{st}^{BP}(x_s, x_t) - c\Big[C_{ts}(x_s; \tau) + C_{st}(x_t; \tau) \Big].$$

The above formulation implies that any zero-gradient must not only satisfy the constraints, but also be of the form defined by BP fixed-point equations.

A.1.2 Gaussian Markov Random Fields

The Gaussian Bethe free energy is formulated in (5.17), and is unbounded below for some models. To enforce boundedness the augmented Lagrangian introduces an additional penalty on log-variance with parameter $\kappa \geq 1$:

$$\mathcal{L}_c(V, \Sigma, \lambda) = \mathcal{F}_{GB}(V, \Sigma) + \sum_s \sum_{t\in N(s)} \lambda_{ts} \left[V_s - V_{ts} \right]$$
$$+ \frac{\kappa}{2} \sum_s \sum_{t\in N(s)} \left[\log V_s - \log V_{ts} \right]^2 + \frac{c}{2} \sum_s \sum_{t\in N(s)} \left[V_s - V_{ts} \right]^2.$$

The derivative w.r.t. the node variance is given by,

$$\frac{\partial \mathcal{L}}{\partial V_s} = \frac{n_s - 1}{2} \left[V_s^{-1} - A_s - \frac{1}{n_s - 1} \sum_{t\in N(s)} \lambda_{st} \right]$$
$$+ c \sum_{t\in N(s)} \left[V_s - V_{ts} \right] + \kappa \sum_{t\in N(s)} \left[\log V_s - \log V_{ts} \right] V_s^{-1}. \tag{A.3}$$

Let the pairwise marginal covariance take the form $\Sigma_{st} = \left(\begin{smallmatrix} V_{ts} & P_{st} \\ P_{ts} & V_{st} \end{smallmatrix}\right)$. Component derivatives of the pairwise covariance are then,

$$\frac{\partial \mathcal{L}}{\partial V_{ts}} = \frac{1}{2}\left[A_s + \lambda_{st} - |\Sigma_{st}|^{-1} V_{st}\right] + c[V_{ts} - V_s] + \kappa[\log V_{ts} - \log V_s] \tag{A.4}$$

$$\frac{\partial \mathcal{L}}{\partial P_{st}} = J_{st} + |\Sigma_{st}|^{-1} P_{st}. \tag{A.5}$$

A.1.3 Discrete Mixtures of Gaussian Potentials

The full joint distribution of the *clutter* model [114] is,

$$\begin{aligned} p(x, z) &= \varphi_0(x) \prod_{i=1}^{n} \psi_0(z_i) \varphi_i(x, z_i; y_i) \\ &= N(x \mid \mu_0, P_0) \prod_{i=1}^{n} (1 - \beta_0)^{1 - z_i} \beta_0^{z_i} N(y_i \mid 0, \sigma_0^2)^{1 - z_i} N(y_i \mid x, \sigma_1^2)^{z_i}. \end{aligned} \tag{A.6}$$

Here β_0 is the Bernoulli prior mean parameter, $x \in \mathbb{R}$ the continuous latent state, $z_i \in \{0, 1\}$ is the discrete clutter indicator, and observations $y_i \in \mathbb{R}$. Using the chain rule for entropy $H(X, Z) = H(Z) + H(X \mid Z)$ we compute the (negative) Bethe entropy as,

$$-H(X, Z) = -\sum_{i=1}^{n} (H(Z_i) + H(X \mid Z_i)) \tag{A.7}$$

$$= \sum_i ((1 - \beta_i) \log(1 - \beta_i) + \beta_i \log \beta_i) - \sum_i ((1 - \beta_i) \frac{1}{2} \log 2\pi e V_{i0} + \beta_i \frac{1}{2} \log 2\pi e V_{i1})$$

Let the marginal approximations be conditional Gaussian q_i and Gaussian q_0 distributions. The Bethe free energy is given by,

$$\mathcal{F}_{CGB}(m, V, \beta) = \sum_{i=1}^{n} \mathbb{E}_i[\log q_i(x, z_i) - \log \phi_i(x, z_i)] - (n - 1)\mathbb{E}_i[\log q_0(x) - \log \varphi_0(x)],$$

Expanding terms:

$$\mathcal{F}_{CGB}(m, V, \beta) = (N - 1)\frac{1}{2} \log V_0 - (N - 1)\frac{1}{2}(V_0 + m_0^2) P_0^{-1} + (N - 1) m_0 P_0^{-1} \mu_0 \tag{A.8}$$

$$\sum_i (1 - \beta_i) \left\{ \log(1 - \beta_i) - \frac{1}{2} \log V_{i0} - \gamma_{i0} + \frac{1}{2}(V_{i0} + m_{i0}^2) P_0^{-1} - m_{i0} P_0^{-1} \mu_0 - \log(1 - \beta_0) \right\} +$$

$$\sum_i \beta_i \left\{ \log \beta_i - \frac{1}{2} \log V_{i1} - \gamma_{i1} + \frac{1}{2}(V_{i1} + m_{i1}^2)(P_0^{-1} + \sigma_1^{-2}) - m_{i1}(P_0^{-1} \mu_0 + \sigma_1^{-2} y_i) - \log \beta_0 \right\}$$

with the shorthand notation $\phi_i(x, z_i) = \varphi_0(x)\psi_0(z_i)\varphi_i(x, z; y_i)$ and $\gamma_{ij} = \log N(y_i \mid 0, \sigma_j^2)$. Note that while the free energy is bounded on the set of expectation constraints [114] the entropy term $\log V_0$ means that the free energy is unbounded below off of the constraint set as $V_0 \to \infty$ at an exponential rate. Such an objective can be problematic for MoM optimization and so we add an additional penalty,

$$\mathcal{F}_{CGB}(m, V, \beta) + \frac{\kappa}{2} \sum_i |\log V_0 - \log \bar{V}_i|^2,$$

for some fixed $\kappa \geq 1$ where the Gaussian mixture variance is denoted,

$$\begin{aligned}
\bar{V}_i &= (1 - \beta_i)V_{i0} + \beta_i V_{i1} + (1 - \beta_i)(m_{i0} - \bar{m}_i)^2 + \beta_i(m_{i1} - \bar{m}_i)^2 \\
\bar{m}_i &= (1 - \beta_i)m_{i0} + \beta_i m_{i1}.
\end{aligned}$$

This added term is quadratic in $\log V_0$, thus bounding the objective off of the constraint set. The augmented Lagrangian is,

$$\begin{aligned}
\mathcal{L}_c(m, V, \beta) = \mathcal{F}(m, V, \beta) &+ \frac{\kappa}{2} \sum_i [\log V_0 - \log \bar{V}_i]^2 + \sum_i \eta_i [m_0 - \bar{m}_i] + \sum_i \lambda_i [V_0 - \bar{V}_i] \\
&+ \frac{c}{2} \sum_i [m_0 - \bar{m}_i]^2 + \frac{c}{2} \sum_i [V_0 - \bar{V}_i]^2
\end{aligned}$$

Gradients of the Gaussian marginal moments are,

$$\begin{aligned}
\frac{\partial \mathcal{L}_c}{\partial V_0} &= (N - 1)\frac{1}{2}V_0^{-1} - (N - 1)\frac{1}{2}P_0^{-1} + \sum_i \lambda_i + c \sum_i [V_0 - \bar{V}_i] + \kappa V_0^{-1} \sum_i [\log V_0 - \log \bar{V}_i] \\
\frac{\partial \mathcal{L}_c}{\partial m_0} &= -(N - 1)m_o P_0^{-1} + (N - 1)P_0^{-1}\mu_0 + \sum_i \eta_i + c \sum_i (m_0 - \bar{m}_i).
\end{aligned}$$

Gradients of the mixture variances,

$$\begin{aligned}
\frac{\partial \mathcal{L}_c}{\partial V_{i0}} &= (1 - \beta_i) \left\{ \frac{1}{2}P_0^{-1} - \frac{1}{2}V_{i0}^{-1} - \lambda_i - c(V_0 - \bar{V}_i) - \kappa(\log V_0 - \log \bar{V}_i)\bar{V}_i^{-1} \right\} \\
\frac{\partial \mathcal{L}_c}{\partial V_{i1}} &= \beta_i \left\{ \frac{1}{2}(P_0^{-1} + \sigma_1^{-2}) - \frac{1}{2}V_{i1}^{-1} - \lambda_i - c(V_0 - \bar{V}_i) - \kappa(\log V_0 - \log \bar{V}_i)\bar{V}_i^{-1} \right\}.
\end{aligned}$$

Gradients of the mixture means,

$$\begin{aligned}
\frac{\partial \mathcal{L}_c}{\partial m_{i0}} = (1 - \beta_i) \Big\{ & m_{i0}P_0^{-1} - P_0^{-1}\mu_0 - \eta_i - c(m_0 - \bar{m}_i) \\
& + 2\beta_i(m_{i1} - m_{i0}) \left[\lambda_i + c(V_0 - \bar{V}_i) + \kappa(\log V_0 - \log \bar{V}_i)\bar{V}_i^{-1} \right] \Big\} \\
\frac{\partial \mathcal{L}_c}{\partial m_{i1}} = \beta_i \Big\{ & m_{i1}(P_0^{-1} + \sigma_1^{-2}) - P_0^{-1}\mu_0 - \sigma_1^{-2} y_i - \eta_i - c(m_0 - \bar{m}_i) \\
& 2(1 - \beta_i)(m_{i0} - m_{i1}) \left[\lambda_i + c(V_0 - \bar{V}_i) + \kappa(\log V_0 - \log \bar{V}_i)\bar{V}_i^{-1} \right] \Big\}
\end{aligned}$$

For the mixture weights we first introduce some shorthand notation,

$$\xi_{i0}(m,V,\beta) = \log(1-\beta_i) - \frac{1}{2}\log V_{i0} - \gamma_{i0} + \frac{1}{2}(V_{i0}+m_{i0}^2)P_0^{-1} - m_{i0}P_0^{-1}\mu_0$$
$$\xi_{i1}(m,V,\beta) = \log\beta_i - \frac{1}{2}\log V_{i1} - \gamma_{i1} + \frac{1}{2}(V_{i1}+m_{i1}^2)(P_0^{-1}+\sigma_1^{-2}) - m_{i1}(P_0^{-1}\mu_0+\sigma_1^{-1}y_i),$$

we similarly define shorthand for partials of the mean and variance constraints,

$$\begin{aligned} m' = \frac{\partial C_i^{mean}}{\partial \beta_i} &= m_{i0} - m_{i1} \\ V' = \frac{\partial C_i^{var}}{\partial \beta_i} &= V_{i0} - V_{i1} + (m_{i0}-\bar{m}_i)^2 - (m_{i1}-\bar{m}_i)^2 \\ &\quad - 2*(m_{i0}-m_{i1})((1-\beta_i)(m_{i0}-\bar{m}) + \beta_i(m_{i1}-\bar{m}_i)) \end{aligned}$$

and the derivative w.r.t. the mixture weights is given by,

$$\begin{aligned} \frac{\partial \mathcal{L}_c}{\partial \beta_i} &= -\xi_{i0}(m,V,\beta) + \xi_{i1}(m,V,\beta) \\ &\quad + m'(\eta_i + c(m_0-\bar{m}_i)) + V'(\lambda_i + c(V_0-\bar{V}_i) + c\bar{V}_i^{-1}(\log V_0 - \log \bar{V}_i)) \end{aligned}$$

A.2 Diverse Particle Selection Proofs

The sections below provide detailed proofs for propositions in Chapter 3. We first show that diverse particle selection of Sec. 3.2 corresponds to a monotonic submodular maximization subject to cardinality constraints. We then shown that the sum of reweighted message differences provides an upper bound on pseudo-max-marginal distortion.

A.2.1 Proof of Prop. 3.2.2

Recall that the message vector over any particle subset is $\hat{m}_{ts}(z)$, where the indicator vector $z \in \{0,1\}^{\alpha N}$ controls particle selection. Recall from Section 3.2 that particles are selected for node $t \in \mathcal{V}$ to minimize total error w.r.t. the augmented message vector m_{ts}, given by the IP:

$$\begin{aligned} &\underset{z}{\text{minimize}} \sum_{s\in\Gamma(t)} \sum_{a=1}^{\alpha N} \left(m_{ts}(a)^{\rho_{ts}} - \hat{m}_{ts}(a;z)^{\rho_{ts}} \right) \\ &\text{subject to} \quad \|z\|_1 \le N, \quad z \in \{0,1\}^{\alpha N}. \end{aligned} \tag{A.9}$$

We drop subscripts and constant terms to yield an equivalent maximization:

$$\underset{z:\|z\|_1\leq N}{\text{maximize}} \sum_a F_a(z) = \sum_a \left[\max_{1\leq b\leq N} z(b)M(a,b)\right]. \tag{A.10}$$

Note that we have dropped subscripts to simplify notation, and the *message foundation matrix* is a compact representation of quantities involved in message calculations,

$$M(a,b) = \psi_t(x_t^{(b)})\psi_{st}(x_s^{(a)}, x_t^{(b)})^{\frac{1}{\rho_{st}}} \frac{\prod_{u\in\Gamma(t)\setminus s} m_{ut}(b)^{\rho_{ut}}}{m_{st}(b)^{1-\rho_{st}}}.$$

Let $y, z \in \{0,1\}^{\alpha N}$ be particle selections and $y \subseteq z$ such that $(y(b)=1) \Rightarrow (z(b)=1)$. For some candidate particle $\overline{b}$:

$$\overline{y}(b) = \begin{cases} 1, & \text{if } b=\overline{b} \\ y(b), & \text{o.w.} \end{cases} \quad \overline{z}(b) = \begin{cases} 1, & \text{if } b=\overline{b} \\ z(b), & \text{o.w.} \end{cases}$$

The margins are given by direct calculation:

$$F_a(\overline{y}) - F_a(y) = \max(0, M(a,\overline{b}) - \hat{m}(a;y))$$
$$F_a(\overline{z}) - F_a(z) = \max(0, M(a,\overline{b}) - \hat{m}(a;z)).$$

Since $y \subseteq z$ we have that F_a is submodular,

$$F_a(\overline{y}) - F_a(y) \geq F_a(\overline{z}) - F_a(z).$$

A sum of submodular functions is submodular, and monotonicity holds since $\hat{m}(y) \leq \hat{m}(z)$.

A.2.2 Proof of Prop. 3.2.1

To simplify we ignore normalization terms and drop dependence on z so $\hat{m}(z) = \hat{m}$. The proof is by induction on the number of neighbors, for the base case let $\Gamma(s) = \{i,j\}$:

$$\begin{aligned}
&\|\nu_s - \hat{\nu}_s\|_1 \leq \dots \\
&\quad \sum_{x_s} \left[(m_{is}(x_s)^{\rho_{is}} - \hat{m}_{is}(x_s)^{\rho_{is}})m_{js}(x_s)^{\rho_{js}} + (m_{js}(x_s)^{\rho_{js}} - \hat{m}_{js}(x_s)^{\rho_{js}})\hat{m}_{is}(x_s)^{\rho_{is}}\right] \\
&\quad \leq \sum_{x_s} \left[(m_{is}(x_s)^{\rho_{is}} - \hat{m}_{is}(x_s)^{\rho_{is}}) + (m_{js}(x_s)^{\rho_{js}} - \hat{m}_{js}(x_s)^{\rho_{js}})\right]
\end{aligned}$$

The first inequality drops $\psi_s \in [0,1]$, and $|\cdot|$ since $\hat{m}_{ts} \preceq m_{ts}$, and the second inequality holds since $m, \hat{m} \in [0,1]$. For the inductive step let $\Gamma(s) = \{t_1, \dots, t_n\}$ and assume:

$$\|\nu_s^{\setminus n} - \hat{\nu}_s^{\setminus n}\|_1 \leq \sum_{i\neq n} \|m_{t_i s} - \hat{m}_{t_i s}\|_1^{\rho_{t_i s}}$$

where $\nu_s^{\setminus n}(x_s)$ is the product of all messages except $m_{t_n s}$:

$$\begin{aligned}\|\nu_s - \hat{\nu}_s\|_1 &\leq \|m_{t_n s} - \hat{m}_{t_n s}\|_1^{\rho_{t_n s}} + \|\nu_s^{\setminus n} - \hat{\nu}_s^{\setminus n}\|_1 \\ &\leq \sum_{i=1}^{n} \|m_{t_i s} - \hat{m}_{t_i s}\|_1^{\rho_{t_i s}}.\end{aligned}$$

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE PAMI*, 34(11):2274–2282, 2012.

[2] S. M. Aji and R. J. McEliece. The generalized distributive law. *Information Theory, IEEE Transactions on*, 46(2):325–343, 2000.

[3] S. Amari. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, pages 357–385, 1982.

[4] S Amari and H Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000.

[5] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice Hall, New Jersey, 1979.

[6] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.

[7] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342, 2010.

[8] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1014–1021. IEEE, 2009.

[9] C. B. Anfinsen. The formation and stabilization of protein structure. *Biochemical Journal*, 128(4):737, 1972.

[10] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, 2011.

[11] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan. *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. John Wiley and Sons, Inc., 2002.

[12] Y. Bar-Shalom, P. K. Willett, and X. Tian. Tracking and data fusion. *A Handbook of Algorithms. Yaakov Bar-Shalom*, 2011.

[13] D. Barber. Expectation correction for smoothed inference in switching linear dynamical systems. *The Journal of Machine Learning Research*, 7:2515–2540, 2006.

[14] D. Barber, A. T. Cemgil, and S. Chiappa. *BAyesian time series models*. Cambridge University Press, 2011.

[15] A. Barbu and S. Zhu. Generalizing swendsen-wang to sampling arbitrary posterior probabilities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1239–1253, 2005.

[16] O. Barndorff-Nielsen. *Information and exponential families in statistical theory*. John Wiley & Sons, 2014.

[17] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse M-best solutions in Markov random fields. In *Computer Vision–ECCV 2012*, pages 1–16. Springer, 2012.

[18] D. S. Berkholz, M. V. Shapovalov, R. L. Dunbrack, and P. A. Karplus. Conformation dependence of backbone geometry in proteins. *Structure*, 17(10):1316–1325, 2009.

[19] C. Berrou. The ten-year-old turbo codes are entering into service. *IEEE Communications magazine*, 41(8):110–116, 2003.

[20] C. Berrou and A. Glavieux. Near optimum error correcting coding and decoding: Turbo-codes. *Communications, IEEE Transactions on*, 44(10):1261–1271, 1996.

[21] D.P. Bertsekas. Constrained optimization and Lagrange multiplier methods. *Computer Science and Applied Mathematics, Boston: Academic Press, 1982*, 1, 1982.

[22] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[23] F. Besse, C. Rother, A. Fitzgibbon, and J. Kautz. PMBP: Patchmatch belief propagation for correspondence field estimation. In *BMVC*, 2012.

[24] S. Bhatia, L. Sigal, M. Isard, and M. J. Black. 3D human limb detection using space carving and multi-view eigen models. In *Computer Vision and Pattern Recognition Wworkshop (CVPR), IEEE Conference on*, pages 17–17. IEEE, 2004.

[25] M. J. Bower, F. E. Cohen, and R. L. Dunbrack Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *Journal of molecular biology*, 267(5):1268–1282, 1997.

[26] K. J. Bowers, E. Chow, H. Xu, R. O. Dror, M. P. Eastwood, B. Gregersen, J. L. Klepeis, I. Kolossvary, M. Moraes, and F. D. Sacerdoti. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC 2006 Conference, Proceedings of the ACM/IEEE*, pages 43–43. IEEE, 2006.

[27] A. Brockwell, P. Del Moral, and A. Doucet. Sequentially interacting Markov chain Monte Carlo methods. *The Annals of Statistics*, 38(6):3387–3411, 2010.

[28] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Structure, Function, and Bioinformatics*, 21(3):167–195, 1995.

[29] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein science*, 12(9):2001–2014, 2003.

[30] O. Cappé, S. J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential monte carlo. *Proceedings of the IEEE*, 95(5):899–924, 2007.

[31] E. Charniak and M. Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics, 2005.

[32] M. Chertkov and V. Y. Chernyak. Loop calculus helps to improve belief propagation and linear programming decodings of low-density-parity-check codes. *Proceedings of the Allerton Conference on Control, Communication and Computing*, 2007.

[33] S. Chung, G. D. Forney, T. J. Richardson, and R. Urbanke. On the design of low-density parity-check codes within 0.0045 db of the Shannon limit. *Communications Letters, IEEE*, 5(2):58–60, 2001.

[34] M. Collins and T. Koo. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70, 2005.

[35] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, and Foldit players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

[36] G. Cornuejols, M. L. Fisher, and G. L. Nemhauser. Exceptional paper-location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management science*, 23(8):789–810, 1977.

[37] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. Probabilistic networks and expert systems. *Statistics for Engineering and Information Science*, 1999.

[38] B. Cseke and T. Heskes. Properties of bethe free energies and message passing in Gaussian models. *Journal of Artificial Intelligence Research*, 41(2):1–24, 2011.

[39] I. Csisz and G. Tusnády. Information geometry and alternating minimization procedures. *Recent results in estimation theory and related topics*, 1984.

[40] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[41] A. P. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2(1):25–26, 1992.

[42] Johan Desmet, Marc De Maeyer, Bart Hazes, and Ignace Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, pages 539–42, 1992.

[43] H. Di, Q. Shi, F. Lv, M. Qin, and Y. Lu. Contour flow: Middle-level motion estimation by combining motion segmentation and contour alignment. In *International Conference on Computer Vision (ICCV)*, 2015.

[44] P. L. Dobruschin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Prob. Appl.*, 13(2):197–224, 1968.

[45] A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo methods in practice*. Springer Verlag, 2001.

[46] R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, 6(8):1661–1681, 1997.

[47] Roland L Dunbrack and Martin Karplus. Backbone-dependent rotamer library for proteins application to side-chain prediction. *Journal of molecular biology*, 230(2):543–574, 1993.

[48] Marcin Eichner and Vittorio Ferrari. Better appearance models for pictorial structures. *British machine vision conference (BMVC)*, 2009.

[49] R. A. Engh and R. Huber. Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica*, 47(4):392–400, 1991.

[50] P. F. Felzenszwalb and D. P Huttenlocher. Efficient matching of pictorial structures. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 66–73. IEEE, 2000.

[51] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[52] Vittorio Ferrari, Manual Marin-Jimenez, and Andrew Zisserman. Progressive search space reduction for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[53] Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, 100(1):67–92, 1973.

[54] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[55] O. Freifeld, A. Weiss, S. Zuffi, and M. J. Black. Contour people: A parameterized model of 2d articulated human shape. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 639–646. IEEE, 2010.

[56] B. J. Frey and D. J. C. MacKay. A revolution: Belief propagation in graphs with cycles. *Advances in neural information processing systems*, pages 479–485, 1998.

[57] M. Fromer and C. Yanover. Accurate prediction for atomic-level protein design and its application in diversifying the near-optimal sequence space. *Proteins: Structure, Function, and Bioinformatics*, 75(3):682–705, 2009.

[58] M. Fromer, C. Yanover, A. Harel, O. Shachar, Y. Weiss, and M. Linial. Sprint: side-chain prediction inference toolbox for multistate protein design. *Bioinformatics*, 26(19):2466–2467, 2010.

[59] Menachem Fromer and Amir Globerson. An lp view of the m-best map problem. *Advances in Neural Information Processing Systems*, 22:567–575, 2009.

[60] R. G. Gallager. *Low-Density Parity-Check Codes*. PhD thesis, MIT Press, Cambridge, 1963.

[61] Dariu M Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[62] I. M. Gelfand and S. V. Fomin. *Calculus of variations*. Courier Dover Publications, 2000.

[63] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *PAMI*, pages 721–741, 1984.

[64] John Geweke. Bayesian inference in econometric models using monte carlo integration. *Econometrica: Journal of the Econometric Society*, pages 1317–1339, 1989.

[65] L. S. Ghoraie, F. Burkowski, S. C. Li, and M. Zhu. Residue-specific side-chain polymorphisms via particle belief propagation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(1):33–41, 2014.

[66] K. Gimpel, D. Batra, C. Dyer, and G. Shakhnarovich. A systematic exploration of diversity in machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

[67] F. Glover and G. A. Kochenberger. *Handbook of metaheuristics*. Springer Science & Business Media, 2003.

[68] Robert F Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophysical journal*, 66(5):1335, 1994.

[69] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993.

[70] V. Granville, M. Křivánek, and J. Rasson. Simulated annealing: A proof of convergence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(6):652–656, 1994.

[71] M. Hassner and J. Sklansky. The use of Markov random fields as models of texture. *Computer Graphics and Image Processing*, 12(4):357–370, 1980.

[72] T. Heskes. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16(11):2379–2413, 2004.

[73] T. Heskes and O. Zoeter. Expectation Propagation for approximate inference in dynamic Bayesian networks. *Uncertainty in Artificial Intelligence*, 18:216–223, 2002.

[74] Tom Heskes, Wim Wiegerinck, Ole Winther, and Onno Zoeter. Approximate inference techniques with expectation constraints. *Journal of Statistical Mechanics: Theory and Experiment*, page 11015, 2005.

[75] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1(1):5–20, 1983.

[76] Lisa Holm and Chris Sander. Fast and simple monte carlo algorithm for side chain optimization in proteins: Application to model building by homology. *Proteins: Structure, Function, and Genetics*, 14(2):213–223, 1992.

[77] H. H Hoos and T. Stützle. *Stochastic local search: Foundations & applications*. Elsevier, 2004.

[78] A. Ihler, J. W. Fisher, and A. S. Willsky. Loopy belief propagation: Convergence and effects of message errors. In *Journal of Machine Learning Research*, pages 905–936, 2005.

[79] A. Ihler and D. McAllester. Particle belief propagation. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS) 2009*, pages 256–263, Clearwater Beach, Florida, 2009. JMLR: W&CP 5.

[80] A. T. Ihler, J. W. Fisher, R. L. Moses, and A. S. Willsky. Nonparametric belief propagation for self-localization of sensor networks. *Selected Areas in Communications, IEEE Journal on*, 23(4):809–819, 2005.

[81] A. T. Ihler, E. B. Sudderth, W. T. Freeman, and A. S. Willsky. Efficient multiscale sampling from products of Gaussian mixtures. *Advances in Neural Information Processing Systems*, 16:1–8, 2003.

[82] Michael Isard. Pampas: Real-valued graphical models for computer vision. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–613. IEEE, 2003.

[83] Michael Isard and Andrew Blake. Condensation - conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[84] E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[85] T. S. Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, page 129, 2001.

[86] T. Joachims, T. Finley, and C. J. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[87] J. E. Jones. On the determination of molecular fields. ii. from the equation of state of a gas. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 106, pages 463–477. The Royal Society, 1924.

[88] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[89] Shanon X Ju, Michael J Black, and Yaser Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 38–44. IEEE, 1996.

[90] K. Kanazawa, D. Koller, and S. Russell. Stochastic simulation algorithms for dynamic probabilistic networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 346–351. Morgan Kaufmann Publishers Inc., 1995.

[91] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, and Z. Popović. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology*, 18(10):1175–1177, 2011.

[92] D. Koller, U. Lerner, and D. Angelov. A general algorithm for approximate inference and its application to hybrid Bayes nets. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 324–333. Morgan Kaufmann Publishers Inc., 1999.

[93] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1568–1583, 2006.

[94] V. Kolmogorov and M. Wainwright. On the optimality of tree-reweighted max-product message-passing. *UAI*, 2005.

[95] R. Kothapa, J. Pacheco, and E. Sudderth. Max-product particle belief propagation. Master's thesis, Brown University, 2011.

[96] F. R. Kschischang, B. J. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498–519, 2001.

[97] B. Kuhlman and D. Baker. Native protein sequences are close to optimal for their structures. *Proceedings of the National Academy of Sciences*, 97(19):10383–10388, 2000.

[98] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *International Conference on Machine Learning (ICML)*, pages 53–61, 2013.

[99] L. Ladický, P. H. S. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR*, pages 3578–3585. IEEE, 2013.

[100] P. T. Lang, H. Ng, J. S. Fraser, J. E. Corn, N. Echols, M. Sales, J. M. Holton, and T. Alber. Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Science*, 19(7):1420–1431, 2010.

[101] S. L. Lauritzen. *Graphical models*. Oxford University Press, 1996.

[102] E. L. Lawler. A procedure for computing the k best solutions to discrete optimization problems and its application to the shortest path problem. *Management science*, 18(7):401–405, 1972.

[103] R. Le Bras, H. Swanger, T. Sereno, G. Beall, and R. Jenkins. Global association final report. Technical report, DTIC Document, 1994.

[104] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the a* algorithm. *Proteins Structure Function and Genetics*, 33(2):227–239, 1998.

[105] C. Lee and S. Subbiah. Prediction of protein side-chain conformation by packing optimization. *Journal of molecular biology*, 217(2):373–388, 1991.

[106] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, pages 420–429. ACM, 2007.

[107] Z. Li and H. A. Scheraga. Monte carlo-minimization approach to the multiple-minima problem in protein folding. *Proceedings of the National Academy of Sciences*, 84(19):6611–6615, 1987.

[108] Simon C Lovell, J Michael Word, Jane S Richardson, and David C Richardson. The penultimate rotamer library. *Proteins: Structure, Function, and Bioinformatics*, 40(3):389–408, 2000.

[109] B. Ma, S. Kumar, C.-J. Tsai, and R. Nussinov. Folding funnels and binding mechanisms. *Protein Engineering*, 12(9):713–720, 1999.

[110] D. J. C. MacKay. Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer, 1998.

[111] Dmitry M. Malioutov, Jason K. Johnson, and Alan S. Willsky. Walk-sums and Belief Propagation in Gaussian graphical models. *Journal of Machine Learning Research*, 7:2031–2064, 2006.

[112] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng. Turbo decoding as an instance of pearl's belief propagation algorithm. *Selected Areas in Communications, IEEE Journal on*, 16(2):140–152, 1998.

[113] T. Minka. The EP energy function and minimization schemes. Technical report, MIT Media Lab, 2001.

[114] T. P. Minka. Expectation Propagation for approximate Bayesian inference. *Uncertainty in Artificial Intelligence*, 17:362–369, 2001.

[115] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, pages 234–243. Springer, 1978.

[116] C. Moallemi and B. Van Roy. Convergence of min-sum message passing for quadratic optimization. *IEEE Transactions on Information Theory*, 55(5):2413–2423, 2009.

[117] Joris M. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, August 2010.

[118] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, and A. Tramontano. Critical assessment of methods of protein structure prediction (CASP)round x. *Proteins: Structure, Function, and Bioinformatics*, 82(S2):1–6, 2014.

[119] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

[120] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Math. Prog.*, 14(1):265–294, 1978.

[121] D. Nilsson. An efficient algorithm for finding the M most probable configurationsin probabilistic expert systems. *Statistics and Computing*, 8(2):159–173, 1998.

[122] J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2(6):522–536, 1980.

[123] J. Pacheco, S. Zuffi, M. Black, and E. Sudderth. Preserving modes and messages via diverse particle selection. In *International Conference on Machine Learning*, pages 1152–1160, 2014.

[124] J. L. Pacheco and E. B. Sudderth. Improved variational inference for tracking in clutter. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 852–855. IEEE, 2012.

[125] B. Paige, F. Wood, A. Doucet, and Y. W. Teh. Asynchronous anytime sequential monte carlo. In *Advances in Neural Information Processing Systems*, pages 3410–3418, 2014.

[126] D. Park and D. Ramanan. N-best maximal decoders for part models. In *ICCV*, pages 2627–2634. IEEE, 2011.

[127] J. Pearl. *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference.* Morgan Kaufmann, San Francisco, CA, 1988.

[128] J. Peng, T. Hazan, D. McAllester, and R. Urtasun. Convex max-product algorithms for continuous MRFs with applications to protein folding. In *Proc. ICML*, 2011.

[129] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[130] Jay W Ponder and Frederic M Richards. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of molecular biology*, 193(4):775–791, 1987.

[131] V. Premachandran, D. Tarlow, and D. Batra. Empirical minimum bayes risk prediction: How to extract an extra few% performance from vision models with just three more parameters. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1043–1050. IEEE, 2014.

[132] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[133] Deva Ramanan and David A Forsyth. Finding and tracking people from the bottom up. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, volume 2, pages II–467. IEEE, 2003.

[134] Deva Ramanan, David A Forsyth, and Andrew Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, volume 1, pages 271–278. IEEE, 2005.

[135] C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

[136] C. Rasmussen. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[137] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using Rosetta. *Methods in enzymology*, 383:66–93, 2004.

[138] E. Rollon, N. Flerova, and R. Dechter. Inference schemes for m best solutions for soft csps. In *Proceedings of Workshop on Preferences and Soft Constraints*, volume 2, 2011.

[139] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1281–1288. IEEE, 2011.

[140] M Schmidt. UGM: A matlab toolbox for probabilistic undirected graphical models. `http://www.cs.ubc.ca/~schmidtm/Software/UGM.html`, 2007.

[141] M. Schmidt, E. Van Den Berg, M. Friedlander, and K. Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm. In *AI & Statistics*, 2009.

[142] M. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 2008.

[143] D. J. Selkoe. Folding proteins in fatal ways. *Nature*, 426(6968):900–904, 2003.

[144] B. Seroussi and J. Golmard. An algorithm directly finding the K most probable configurations in bayesian networks. *International Journal of Approximate Reasoning*, 11(3):205–233, 1994.

[145] E. Shakhnovich. Protein folding thermodynamics and dynamics: where physics, chemistry, and biology meet. *Chemical reviews*, 106(5):1559–1588, 2006.

[146] M. V. Shapovalov and R. L. Dunbrack. Statistical and conformational analysis of the electron density of protein side chains. *Proteins: Struct., Func., and Bioinf.*, 66(2):279–303, 2007.

[147] M. V. Shapovalov and R. L. Dunbrack. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011.

[148] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. Bank, C. Young, M. M. Deneroff, B. Batson, and K. J. Bowers. Millisecond-scale molecular dynamics simulations on anton. In *High Performance Computing Networking, Storage and Analysis, Proceedings of the Conference on*, pages 1–11. IEEE, 2009.

[149] S. E. Shimony. Finding maps for belief networks is NP-hard. *Artificial Intelligence*, 68(2):399–410, 1994.

[150] L. Sigal, A. O Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4–27, 2010.

[151] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *International journal of computer vision*, 98(1):15–48, 2012.

[152] J. Snoek, H. Larochelle, and R.P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[153] D. Sontag, T. Meltzer, A. Globerson, Y. Weiss, and T. Jaakkola. Tightening LP relaxations for MAP using message passing. *Uncertainty in artificial intelligence (UAI)*, 24, 2008.

[154] Stergios Stergiopoulos. *Advanced signal processing handbook: theory and implementation for radar, sonar, and medical imaging real time systems*. CRC press, 2010.

[155] R. L. Streit and T. E. Luginbuhl. Probabilistic multi-hypothesis tracking. Technical Report 10,428, Naval Undersea Warfare Center, Division Newport Rhode Island, Feb. 1995.

[156] E. B. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology, 2006.

[157] E. B. Sudderth, A. T. Ihler, M. Isard, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. *Communications of the ACM*, 53(10):95–103, 2010.

[158] E. B. Sudderth, M. Mandel, W. T. Freeman, and A. S. Willsky. Visual hand tracking using nonparametric belief propagation. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 189–189. IEEE, 2004.

[159] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *IJCV*, 106(2):115–137, 2014.

[160] B. Kappen T. Heskes, K. Albers. Approximate inference and constrained optimization. *Uncertainty in Artificial Intelligence*, 13:313–320, 2003.

[161] S. C. Tatikonda and M. I. Jordan. Loopy belief propagation and Gibbs measures. In *UAI*, pages 493–500. Morgan Kaufmann Publishers Inc., 2002.

[162] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: how accurate are they? *Journal of molecular biology*, 257(2):457–469, 1996.

[163] H. Trinh and D. McAllester. Unsupervised learning of stereo vision with monocular cues. In *British Machine Vision Conference*, 2009.

[164] Z. Tu and S. Zhu. Image segmentation by data-driven markov chain monte carlo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):657–673, 2002.

[165] M. Vendruscolo. Determination of conformationally heterogeneous states of proteins. *Current opinion in structural biology*, 17(1):15–20, 2007.

[166] M. Wainwright, T. Jaakkola, and A. Willsky. Tree consistency and bounds on the performance of the max-product algorithm and its generalizations. *Statistics and Computing*, 14(2):143–166, 2004.

[167] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree-reweighted Belief Propagation algorithms and approximate ML estimation by pseudo-moment matching. In *In AISTATS*, 2003.

[168] M. J. Wainwright, T. S. Jaakkola, and A. S Willsky. Map estimation via agreement on trees: message-passing and linear programming. *Information Theory, IEEE Transactions on*, 51(11):3697–3717, 2005.

[169] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, UC Berkeley, Dept. of Statistics, 2003.

[170] Y. Wald and A. Globerson. Tightness results for local consistency relaxations in continuous mrfs. In *Proc. 30th Uncertainty in Artifical Intelligence (UAI)*, 2014.

[171] Y. Weiss, C. Yanover, and T. Meltzer. Map estimation, linear programming and belief propagation with convex free energies. In *UAI*, 2007.

[172] M. Welling and Y.W. Teh. Belief optimization for binary networks: A stable alternative to Loopy Belief Propagation. In *Uncertainty in Artificial Intelligence*, 2001.

[173] C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. In *Multimodal Technologies for Perception of Humans*, pages 509–519. Springer, 2008.

[174] J. Xu and B. Berger. Fast and accurate algorithms for protein side-chain packing. *Journal of the ACM (JACM)*, 53(4):533–557, 2006.

[175] P. Yadollahpour, D. Batra, and G. Shakhnarovich. Discriminative re-ranking of diverse segmentations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1923–1930. IEEE, 2013.

[176] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011.

[177] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *IEEE PAMI*, 35(12):2878–2890, December 2013.

[178] C. Yanover, T. Meltzer, and Y. Weis. Linear programming relaxations and belief propagation–an empirical study. *The Journal of Machine Learning Research*, 7:1887–1907, 2006.

[179] C. Yanover and Y. Weiss. Approximate inference and protein-folding. In *Advances in neural information processing systems*, pages 1457–1464, 2002.

[180] C. Yanover and Y. Weiss. Finding the M most probable configurations using loopy belief propagation. *Advances in Neural Information Processing Systems*, 16:289, 2003.

[181] J. S. Yedidia. An idiosyncratic journey beyond mean field theory. advanced mean field methods, theory and practice 21-36, 2001.

[182] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized Belief Propagation. *Advances in neural information processing systems*, pages 689–695, 2001.

[183] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282–2312, 2005.

[184] A. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to Belief Propagation. *Neural Computation*, 14:1691–1722, 2002.

[185] Wojtek Zajdel, Johannes D Krijnders, T Andringa, and Dariu M Gavrila. Cassandra: audio-video sensor fusion for aggression detection. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 200–205. IEEE, 2007.

[186] O. R. Zoeter and T. M. Heskes. Gaussian quadrature based expectation propagation. *Workshop on Artificial Intelligence and Statistics*, 2005.

[187] S. Zuffi, O. Freifeld, and M. Black. From pictorial structures to deformable structures. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3546–3553. IEEE, 2012.

[188] S. Zuffi, J. Romero, C. Schmid, and M. Black. Estimating human pose with flowing puppets. In *International conference on computer vision*, 2013.