Mode-Marginals: Expressing Uncertainty via Diverse M-Best Solutions

Varun Ramakrishna
Robotics Institute
Carnegie Mellon University
varunnr@cmu.edu

Dhruv Batra
Virginia Tech
dbatra@vt.edu

Abstract

Representing uncertainty in predictions made by complex probabilistic models is often crucial but computationally challenging. There are a number of useful probabilistic models where computing the most probably assignment (or MAP) is easy but finding the marginal probability of variables is hard. In this paper, we present a novel representation of uncertainty in discrete probabilistic models, that we call *Mode-Marginals*. Mode-Marginals contain both max-marginals and marginals as special cases and express the entire spectrum in between the two. Computing mode-marginals involves performing MAP computation on *deterministic* perturbations to the unnormalized probability function, as introduced by Batra *et al*. [1]. We evaluate our method on a challenging computer vision application – human pose estimation, and demonstrate that mode-marginals provide a richer representation of uncertainty than max-marginals, leading to improved performance.

1 Introduction

Intelligent systems must operate under significant levels of uncertainty. Real-world data – images from a camera, text from Twitter feeds, output from gene sequencers – are often noisy and ambiguous. Probabilistic models provide a principled framework for dealing with uncertainty and for converting evidence from (multiple) noisy sources into a posteriori *belief* about the world.

Unfortunately, reasoning about the full posterior is typically intractable, resulting a major formal divide – our models can either reason about the full posterior by making performance-limiting independence assumptions, or mismanage uncertainty by modeling and predicting only the most probable or maximum a posteriori (MAP) hypothesis. The former are often too restrictive to capture reality but allow a full probabilistic treatment, *e.g*. computing the *marginal* distribution of a subset of the variables. The latter include more expressive models but do not provide any measure of uncertainty in their prediction, causing errors to propagate to systems that depend on these predictions.

At the heart of this divide lies the fact that there exist useful models where *maximizing* over an unnormalized posterior (or score) is easy, but computing the normalizing constant by *summing* over scores of all states is hard. For instance, a maximum bipartite matching can be found in $O(n^3)$ time with the Hungarian algorithm [5], but summing over all perfect matchings (*i.e*. computing the permanent) is #P-complete [9].

This work is a step towards bridging this divide by computing estimates of uncertainty that do not require summation, rather re-use the techniques used for finding the maximizing assignment. At a high-level, this is similar in spirit to a recent line of work [3, 6, 8].

Max-Marginals and Marginals. One common approach for representing uncertainty in the MAP solution is to use max-marginals (sometimes also referred to as min-marginals). A max-marginal for variable x_i and state s is the score of the highest scoring configuration of the full model, under the constraint that variable x_i is always assigned state s, *i.e*. max-marg$(x_i = s) = \max_{\mathbf{x}, x_i = s} P(\mathbf{x})$. Compare this to a marginal, which sums over all configurations of other variables, *i.e*. marg$(x_i =$

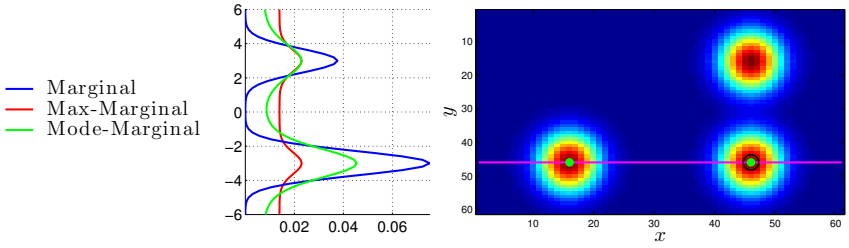

Figure 1: **A Toy Example.** Consider the joint probability distribution of two variables (x and y), with three peaks of equal heights. The marginal, max-marginal and mode-marginals are plotted for variable y. A slice of the joint distribution used to calculate the marginal distribution for y is plotted on the figure as the magenta line. The max-marginal approximates the marginal by taking the maximum scoring sample denoted by the black circle, and thus cannot distinguish between a single peak and two peaks. The mode-marginal samples the two modes (denoted by the green circles) to obtain a more accurate approximation of the marginal.

$s) = \sum_{\mathbf{x}, x_i = s} P(\mathbf{x})$. Intuitively, a max-marginal *approximates* a marginal by replacing a slice of the joint distribution – $P(x_1, \ldots, x_{i-1}, x_i = s, x_{i+1}, \ldots, x_n)$ – by a delta function centered at its mode. When viewed from this perspective, the main motivation of this paper is easy to understand – can we produce better approximations to the marginal while still staying tractable?

Overview. In this paper, we present a novel representation of uncertainty that we call *mode-marginals* (M-MARGS). Intuitively, M-MARGS involve approximating the slice of the joint distribution — $P(x_1, \ldots, x_{i-1}, x_i = s, x_{i+1}, \ldots, x_n)$ — by not just a single delta function centered at its mode, rather a *collection of delta functions*, each centered at a local peak in the slice. Specifically, we leverage the recent work of Batra *et al*. [1] to mine for the diverse M-best joint assignments that fix the variable in question. Ultimately, such a collection of delta functions provides a more accurate approximation to the exponentially large slice. M-MARGS contain max-marginals as a special case, and can be viewed as occupying the spectrum between max-marginals and marginals, controlled by the parameter M. See Figure 1 for an example.

Our experiments show that M-MARGS provide a richer representation of uncertainty than max-marginals, leading to improved performance in challenging application like human-pose estimation where large variation in appearance and articulations result in multi-modal posterior distributions.

2 Preliminaries: MAP and Max-Marginals

Notation. For any positive integer n, let $[n]$ be shorthand for the set $\{1, 2, \ldots, n\}$. We consider a set of discrete random variables $\mathbf{x} = \{x_i \mid i \in [n]\}$, each taking value in a finite label set, $x_i \in X_i$. For a set $A \subseteq [n]$, we use x_A to denote the tuple $\{x_i \mid i \in A\}$, and X_A to be the cartesian product of the individual label spaces $\times_{i \in A} X_i$. For ease of notation, we use x_{ij} as a shorthand for $x_{\{i,j\}}$.

MAP. Let $G = (\mathcal{V}, \mathcal{E})$ be a graph defined over these variables, *i.e.* $\mathcal{V} = [n]$, $\mathcal{E} \subseteq \binom{\mathcal{V}}{2}$. Let $\theta_i(s)$ be the unary term expressing the local confidence at site i for the label s, and $\theta_{ij}(s,t)$ be the pairwise term expressing compatibility of label s and t at adjacent nodes. The *score* for any configuration $\mathbf{x}$ is given by the sum $S(\mathbf{x}) = \sum_{i \in \mathcal{V}} \theta_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(x_i, x_j)$, and its probability is given by the Gibbs distribution: $P(\mathbf{x}) = \frac{1}{\mathcal{Z}} e^{S(\mathbf{x})}$, where $\mathcal{Z}$ is the partition function or the normalization constant.

The goal of MAP inference is to find the most probable joint assignment of variables:

$$\arg\max_{\mathbf{x} \in X_{\mathcal{V}}} P(\mathbf{x}) = \arg\max_{\mathbf{x} \in X_{\mathcal{V}}} \sum_{i \in \mathcal{V}} \theta_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(x_i, x_j). \tag{1}$$

Normalized and Unnormalized Max-Marginals. A max-marginal describes the Gibbs distribution under certain constraints. Specifically, it is the maximum possible probability when the state for a single variable is specified: $\mu_i(s) = \max_{\mathbf{x} \in X_{\mathcal{V}}, x_i = s} P(\mathbf{x})$. In practice, inference algorithms work with the unnormalized distribution (or score) and thus produce an unnormalized max-marginal: $\Phi_i(s) = \max_{\mathbf{x} \in X_{\mathcal{V}}, x_i = s} \exp(S(\mathbf{x}))$. Note that both quantities are equal up to a scaling factor, *i.e.*: $\frac{\mu_i(s)}{\sum_{t \in X_i} \mu_i(t)} = \frac{\Phi_i(s)}{\sum_{t \in X_i} \Phi_i(t)}$.

3 A New Representation: Mode-Marginals

We now introduce a new representation of uncertainty that we call *mode-marginals* (M-MARGS).

M-MARGS for variable x_i and state s is characterized by a set of M configurations such that x_i is assigned state s in all these configurations, $\mathcal{X}_{is} = \{\mathbf{x}^m \mid x_i^m = s \;\; \forall m \in [M]\}$. Formally, M-MARGS is defined as

$$\text{M-MARGS:} \qquad \Phi_i^M(s) = \sum_{\mathbf{x} \in \mathcal{X}_{is}} \exp\left(S(\mathbf{x})\right) \tag{2}$$

We can see that M-MARGS computes a sum of (exponentiated) scores for the entries in $\mathcal{X}_{is}$.

M-MARGS contains both max-marginals and marginals as a special case, and expresses the entire spectrum in between the two. If $\mathcal{X}_{is}$ is the set of *all possible* configurations $\{\mathbf{x} \in X_{\mathcal{V}} \mid x_i = s\}$, then M-MARGS essentially becomes the marginal. If $\mathcal{X}_{is}$ is simply the single best configuration $\operatorname{argmax}_{\mathbf{x} \in X_{\mathcal{V}} | x_i = s} S(\mathbf{x})$, then M-MARGS becomes the max-marginal.

Thus, the set $\mathcal{X}_{is}$ controls how well M-MARGS approximates the marginal. There are two main criteria for choosing elements of $\mathcal{X}_{is}$. These configurations should 1) provide an accurate summary of score and 2) be efficiently computable.

Creating Marginal Summaries with DivMBest Solutions. To satisfy both criteria, we leverage the the recently proposed method of of Batra *et al*. [1], called DivMBest to mine for the diverse M-best joint assignments that fix the variable in question.

DivMBest finds diverse M-best solutions incrementally. Let $\mathbf{x}^1$ be the best solution, $\mathbf{x}^2$ be the second DivMBest solution and so on. At each step, the next best solution is defined as the highest scoring state with a minimum degree of "dissimilarity" w.r.t. previously chosen solutions, where dissimilarity is measured under a function $\Delta(\cdot, \cdot)$:

$$\mathbf{x}^M = \operatorname*{argmax}_{x \in X_{\mathcal{V}}, x_i = s} \sum_{i \in \mathcal{V}} \theta_i(x_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(x_i, x_j) \tag{3}$$

$$s.t. \qquad \Delta(\mathbf{x}, \mathbf{x}^m) \geq k_m \qquad\qquad \forall m \in \{1, \ldots, M-1\}. \tag{4}$$

In general, DivMBest$(\Delta, \boldsymbol{k})$ is at least as hard as the MAP inference problem which is NP-hard. Thus, Batra *et al*. [1] proposed to use the Lagrangian relaxation of DivMBest$(\Delta, \boldsymbol{k})$, formed by *dualizing* the dissimilarity constraints $\Delta(\mathbf{x}, \mathbf{x}^m) \geq k_m$:

$$f(\boldsymbol{\lambda}) = \max_{x \in X_{\mathcal{V}}, x_i = s} S_\Delta(\mathbf{x}) = \max_{x \in X_{\mathcal{V}}, x_i = s} \sum_{A \in \mathcal{V} \cup \mathcal{E}} \theta_A(y_A) + \sum_{m=1}^{M-1} \lambda_m \Big(\Delta(\mathbf{x}, \mathbf{x}^m) - k_m \Big). \tag{5}$$

Here $\boldsymbol{\lambda} = \{\lambda_m \mid m \in [M-1]\}$ is the set of Lagrange multipliers, which determine the weight of the penalty imposed for violating the constraints. Intuitively, we see that the Lagrangian relaxation maximizes a Δ-augmented score, *i.e*., a linear combination of the MRF score and the dissimilarity w.r.t. the previous solutions, with the weighting given by the Lagrange multipliers. For some classes of Δ-functions, we can solve the Δ-augmented score maximization problem using the *same algorithms* used for finding the MAP.

Example: Hamming Dissimilarity. Consider $\Delta(\mathbf{x}, \mathbf{x}^m) = \sum_{i \in \mathcal{V}} [\![x_i \neq x_i^m]\!]$, where $[\![\cdot]\!]$ is the Iverson bracket (1 if input condition is true, 0 otherwise). This function counts the number of nodes that are labeled differently between two solutions. For this dissimilarity function, the Δ-augmented scoring function can be written as:

$$S_\Delta(y) = \sum_{i \in \mathcal{V}} \underbrace{\left(\boldsymbol{\theta}_i(x_i) + \sum_{m=1}^{M-1} \lambda_m [\![x_i \neq x_i^m]\!] \right)}_{\text{Perturbed Unary Score}} + \sum_{(i,j) \in \mathcal{E}} \boldsymbol{\theta}_{ij}(x_i, x_j). \tag{6}$$

This perturbed maximization can be performed via the same algorithm used for MAP inference.

4 Application: Mode-Marginals for Part Proposals in Human Pose

We demonstrate the use of mode-marginals in the task of estimating human pose from a single image, which is a challenging task due to the large variation in configuration and appearance. The

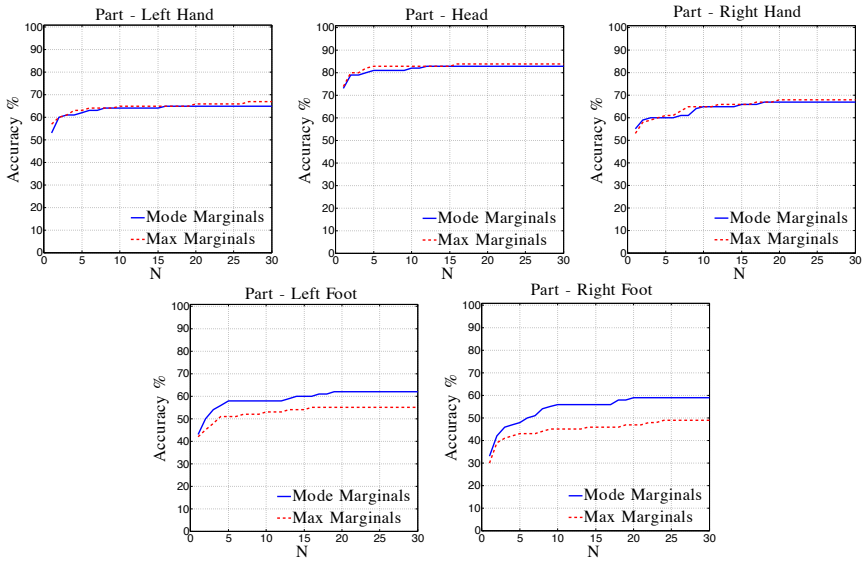

Figure 2: **Part Retrieval Accuracy** for the head, hands and feet on the PARSE test set. We can see that M-MARGS significantly outperforms max-marginals on both feet parts. (N in the case of max-marginals corresponds to generating N proposals from the max-marginal score map by performing non-maxima suppression.)

Figure 3: **Mode-marginal hough score maps.** We show color-coded hough score maps for two example images from the PARSE dataset. The hough map for each body part is color-coded.

pose of people in cluttered images is often ambiguous, leading to multi-modal CRF distributions. We show that M-MARGS help express uncertainty in such multi-modal domains.

Model. We use the articulated part-model of Yang and Ramanan [10], which models human pose in an image as a tree-structured CRF. The location of each part p of the body in the image is a variable $\mathbf{x}_p = (x_p, y_p) \in \mathrm{R}^2$ resulting in a configuration vector given by $\mathbf{x} = (x_1, \ldots x_P) \in \mathrm{R}^{2P}$. We compute M-MARGS for this model with $M = 5$.

Experiments. We evaluate the quality of M-MARGS by using them in a part-retrieval setting, *i.e.* for each image we compute the M-MARGS for the head, hands and feet and take the top N scoring configurations for each part. We then measure the retrieval accuracy at N, and compare M-MARGS to max-marginals with the same set-up. The results are shown in Figure 2. We can see that M-MARGS performs similar to max-marginals for head and hands, while significantly outperforming max-marginals for the feet. The feet tend to be most prone to double-counting[1] errors in tree-structured models for human pose. Our method leverages diverse high scoring configurations and enables us to achieve higher part-retrieval accuracy. Such an improved performance in part-location proposals benefits tracking systems such as [2] that first produce part-location hypotheses and then track parts. The DivMBest configurations at each location of each part can be used to vote for the placement of every other part in a hough-voting scheme. The color-coded hough-score maps are shown in Figure 3.

[1]When the same image evidence is used to explain two parts with the same appearance.

References

[1] D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In *ECCV (5)*, 2012. 1, 2, 3

[2] P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. *International journal of computer vision*, 95(2):180–197, 2011. 5

[3] T. Hazan and T. Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *International Conference of Machine Learning (ICML)*, 2012. 1

[4] P. Kohli and P. H. S. Torr. Measuring uncertainty in graph cut solutions. *Comput. Vis. Image Underst.*, (1), Oct. 2008.

[5] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, 5(1):32–38, March 1957. 1

[6] G. Papandreou and A. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*, pages 193–200, Nov. 2011. 1

[7] D. Tarlow and R. Adams. Revisiting uncertainty in graph cut solutions. In *CVPR*, june 2012.

[8] D. Tarlow, R. P. Adams, and R. S. Zemel. Randomized optimum models for structured prediction. In *Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012. 1

[9] L. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189 – 201, 1979. 1

[10] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011. 4