Learning with Maximum A-Posteriori Perturbation Models

Andreea Gane
CSAIL, MIT

Tamir Hazan
University of Haifa, Israel

Tommi Jaakkola
CSAIL, MIT

Abstract

Perturbation models are families of distributions induced from perturbations. They combine randomization of the parameters with maximization to draw unbiased samples. Unlike Gibbs' distributions, a perturbation model defined on the basis of low order statistics still gives rise to high order dependencies. In this paper, we analyze, extend and seek to estimate such dependencies from data. In particular, we shift the modelling focus from the parameters of the Gibbs' distribution used as a base model to the space of perturbations. We estimate dependent perturbations over the parameters using a hard-EM approach, cast in the form of inverse convex programs. Each inverse program confines the randomization to the parameter polytope responsible for generating the observed answer. We illustrate the method on several computer vision problems.

1 Introduction

Realistic modeling approaches are built on flexible probability distributions and inference algorithms that support them. For example, problems such as scene understanding [5], parsing [21], or protein design [32] all involve complex inference calculations in models where likely structures, parses, or arrangements are guided by potential functions over subsets of variables. By introducing higher order potential functions we obtain richer (more realistic) models but at the cost of heavy inference calculations.

Feasibility of inference calculations is often linked to Markov properties such as conditional independencies. Markov properties are exploited in both exact and approximate inference algorithms, including belief propagation [30], Gibbs sampling [8], Metropolis-Hastings [12] or Swendsen-Wang [39]. In specific cases one can sample efficiently from a Markov random field model by constructing a rapidly mixing Markov chain (cf. [16, 17, 15]). Such approaches do not extend to many practical cases where the values of the variables are strongly guided by both data (high signal) and prior knowledge (high coupling). Indeed, sampling in high-signal high-coupling regime is known to be provably hard [16, 9].

Finding a single most likely assignment (MAP structure) is often considerably easier than summing over the values of variables or drawing an unbiased sample. Substantial effort has gone into developing algorithms for recovering MAP assignments, either based on specific structural restrictions such as super-modularity [20] or by devising linear programming relaxations and successively refining them [32, 40]. MAP inference is nevertheless limiting when there are a number of alternative likely assignments.

Recently, MAP inference has been combined with randomization to define new classes of probability models that are easy to sample from [29, 34, 13, 14, 27, 24]. These *perturbation models* involve randomization of Gibbs' potentials and finding the corresponding maximizing assignment. The assignment represents a sample from the induced distribution. Properties of the induced distribution are heavily governed by the randomization. Indeed, in contrast to Gibbs' distributions, low order potentials, after undergoing randomization and maximization, lead to high order dependencies in the induced distributions. We seek to understand, extend, and exploit such dependencies.

In this paper, we introduce dependent perturbations as a modeling tool. Perturbation models are latent variable models and we learn distributions over perturbations using a hard-EM approach. In the E-step, an inverse convex program is used to confine the randomization to the parameter polytope responsible for generating the observed answer. We illustrate the approach on several computer vision problems.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

2 Background

Consider a complex model guided by real valued finite potentials $\theta(x) = \theta(x_1, ..., x_n)$ over a discrete product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. The domain is implicitly defined through $\theta(x)$ via exclusions $\theta(x) = -\infty$ whenever $x \notin dom(\theta)$. Potentials are mapped to the probability scale via the Gibbs' distribution:

$$p(x_1, ..., x_n) = \frac{1}{Z(\theta)} \exp(\theta(x_1, ..., x_n)) \quad (1)$$

Such distributions are often useful, but are challenging to learn and sample from, depending on how the potential function decomposes.

Our approach is based on representations of the Gibbs' distribution using the statistics of randomized potentials. We add a random function $\gamma : \mathcal{X} \to R$ to the potential function in (1) and draw a sample by solving the resulting MAP prediction problem:

$$x^* = \arg\max_{x \in \mathcal{X}} \{\theta(x) + \gamma(x)\}. \quad (2)$$

The distribution induced in this manner is given by the following:

$$\mathcal{P}(\hat{x}) = P_\gamma \Big[\hat{x} \in \arg\max_{x \in X} \{\theta(x) + \gamma(x)\} \Big] \quad (3)$$

and it's heavily dependent on the nature of randomization. The simplest approach to designing a perturbation function is to associate an i.i.d. random variable $\gamma(x)$ for each $x \in \mathcal{X}$. The following result characterizes the induced distribution in this case, assuming perturbations are Gumbel distributed:

Theorem 1. *[10] Let $\mathcal{X}$ be finite and let $\{\gamma(x), x \in \mathcal{X}\}$ be a collection of i.i.d. zero mean Gumbel distributed random variables, whose cumulative distribution functions is $F(t) = \exp(-\exp(-(t+c)))$ and $c \approx 0.5772$ is the Euler-Mascheroni constant. Then*

$$P_\gamma \Big[\hat{x} \in \arg\max_{x \in \mathcal{X}} \{\theta(x) + \gamma(x)\} \Big] = \frac{1}{Z(\theta)} \exp(\theta(x)) \quad (4)$$

The max-stability of the Gumbel distribution shows that one can preserve the Markov properties of the Gibbs model using high dimensional perturbations.

Clearly, instantiating $\gamma(x)$ for each $x \in \mathcal{X}$ is not feasible in practice. In our work, we investigate low dimensional perturbations as the main tool to control the dependencies of the induced probability model.

3 Dependencies in Tree Structured Models

Gibbs distributions dependency structures follow their Markov blankets. We investigate whether the same is true for perturbation models.

Theorem 1 implies that when the Gibbs distribution is independent it can be represented using low dimensional perturbation models. To verify this property we recall that a probability distribution is independent whenever $p(x) = \prod_{i=1}^n p(x_i)$, where $p(x_i) = \sum_{x \setminus x_i} p(x)$ are its marginal probabilities. Thus this assertion holds when applying Theorem 1 for each dimension $i = 1, ..., n$ while setting $\theta_i(x_i) = \log p(x_i)$ and using i.i.d. perturbations $\gamma_i(x_i)$ that follow the Gumbel distribution.

In the following we consider the different dependencies that relate to tree structured MRFs. We show that there are perturbations that preserve the Markov properties of the Gibbs distributions. We also show that other perturbations demonstrate long-range dependencies beyond those that are described by the tree structure of its potentials. This is a promising result that implies that perturbation models may learn efficiently long-range dependencies that are currently present in many machine learning applications.

3.1 Perturbation Models with Markov-Type Dependencies

Distributions can be described by their conditional probabilities $p(x_1, ..., x_n) = \prod_{j=1}^n p(x_j | x_1, ..., x_{j-1})$, and in Markov random fields these conditional probabilities are simplified by their dependency graphs. Specifically, assume a tree structured MRF and let $\vec{E}$ be any directed version of the tree. For notational convenience, assume that the vertices $\{1, ..., n\}$ are topologically sorted and that there is an arc $(i \to j)$. Then $p(x_j | x_1, ..., x_{j-1}) = p(x_j | x_i)$. Furthermore, for a tree, specifying $\theta(x)$ is equivalent to specifying marginals probabilities $p(x_i)$, $i = 1, \ldots, n$, and $p(x_i, x_j)$, $(i, j) \in E$, which can be related as follows:

$$\theta_i(x_i) = \log p(x_i), \quad \theta_{ij}(x_i, x_j) = \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \quad (5)$$

The following theorem shows that in this case, for any potential function there are low dimensional perturbation models that preserve these the independencies:

Theorem 2. *Consider the Gibbs distribution with a tree structured Markov random field. Then for any potential function*

$$\theta(x) = \sum_{i=1}^n \theta_i(x_i) + \sum_{(i,j) \in E} \theta_{ij}(x_i, x_j) \quad (6)$$

there are random variables $\{\gamma_{ij}(x_i, x_j)\}$ indexed by $(i, j) \in E, (x_i, x_j) \in \mathcal{X}_i \times \mathcal{X}_j$ such that

$$p(\hat{x}) = P_\gamma \Big[\hat{x} \in \arg\max_{x \in \mathcal{X}} \{\theta(x) + \sum_{(i,j) \in E} \gamma_{ij}(x_i, x_j)\} \Big] \quad (7)$$

Proof. Let $\hat{\gamma}_{ij}(x_i, x_j)$ be i.i.d. random variables that follow the Gumbel distributions. Let $\vec{E}$ be a directed version of the tree and assume that the vertices $\{1, ..., n\}$ are topologically sorted and that there is an arc $(1 \to 2)$. Let $\gamma_{12}(x_1, x_2) = \hat{\gamma}_{12}(x_1, x_2)$ and for any other edge $(i \to j)$ define $\gamma_{ij}(x_i, x_j) =$

$$\hat{\gamma}_{ij}(x_i, x_j) - \max_{x'_j} \left\{\theta_{ij}(x_i, x'_j) + \theta_j(x'_j) + \hat{\gamma}_{ij}(x_i, x'_j)\right\}$$

Let $p(x_1, x_2) = \sum_{x \setminus \{x_1, x_2\}} p(x)$ be the marginal probabilities of Gibbs distribution. We begin by showing that

$$p(\hat{x}_1, \hat{x}_2) = P_\gamma\big[\hat{x}_1, \hat{x}_2 \in \arg\max_{x \in \mathcal{X}} \{\theta(x) + \sum_{(i,j) \in \vec{E}} \gamma_{ij}(x_i, x_j)\}\big]$$

To this end, any sample $(\hat{x}_1, \hat{x}_2)$ from the induced marginal distribution is obtained by

$$\begin{aligned} \hat{x}_1, \hat{x}_2 &= \arg\max_{x_1, x_2} \max_{x \setminus \{x_1, x_2\}} \left\{\theta(x) + \sum_{(i,j) \in \vec{E}} \gamma_{ij}(x_i, x_j)\right\} \\ &= \arg\max_{x_1, x_2} \left\{\log p(x_1, x_2) + \gamma_{12}(x_1, x_2)\right\} \end{aligned}$$

where the equality follows from the definition of $\gamma_{ij}(x_i, x_j)$ that enforces $\max_{x_j}\{\theta_{ij}(x_i, x_j) + \theta_j(x_j) + \gamma_{ij}(x_i, x_j)\} = 0$, applied recursively to each leaf in the tree. Theorem 1 implies that marginal probabilities of the Gibbs distribution and the MAP perturbation distribution are the same since $\gamma_{12}(x_1, x_2)$ are independent Gumbel random variables.

To complete the proof we show that for every $(i \to j)$ the conditional probability of MAP perturbations is the same as the Gibbs. For that end, define for every $\alpha \subset \{1, .., n\}$ the subset of indexes $x_\alpha = (x)_{i \in \alpha}$, and

$$\Gamma(\hat{x}_\alpha) = \left\{\gamma : \ \hat{x}_\alpha \in \arg\max_{x \in \mathcal{X}} \{\theta(x) + \sum_{(i,j) \in \vec{E}} \gamma_{ij}(x_i, x_j)\}\right\}$$

Recall the vertices are topologically ordered, thus we aim at showing that

$$p(x_j | x_i) = P_\gamma\Big(\Gamma(x_1, ..., x_j) | \Gamma(x_1, ..., x_{j-1})\Big)$$

By our construction, for any values of $x_1, ..., x_{j-1}$ the argument x_j is chosen to maximize $\theta_j(x_j) + \theta_{ij}(x_i, x_j) + \hat{\gamma}_{ij}(x_i, x_j)$. Since $\theta_j(x_j) + \theta_{ij}(x_i, x_j) = \log p(x_j | x_i)$ and $\hat{\gamma}_{ij}(x_i, x_j)$ are i.i.d. with zero mean Gumbel distribution, the result follows by applying Theorem 1. □

The perturbation models may describe tree structured Gibbs distributions. Perhaps surprisingly, the random variables that enforce the Markov properties in this case are not independent nor identically distributed. This demonstrates the potential power of induced models when allowing dependent perturbation variables. In the following we show that perturbation models may induce long-range dependencies as well.

3.2 Perturbation Models and Long-Range Dependencies

Next we show that independent low dimensional perturbations may capture long-range interactions. We focus on perturbation models with tree structured potential functions and perturbations on the edge potentials, but the results can be generalized to more complex graphs.

The following theorem shows that when i.i.d. perturbations follow the edge structure of the potential function, we are able to capture dependencies above and beyond the initial structure.

Theorem 3. *There exist perturbation models with tree structured potential functions and i.i.d. perturbation variables $\{\gamma_{ij}(x_i, x_j)\}$ indexed by $(i, j) \in E, (x_i, x_j) \in \mathcal{X}_i \times \mathcal{X}_j$, such that the induced model given by:*

$$\mathcal{P}(\hat{x}) = P_\gamma\Big[\hat{x} \in \arg\max_{x \in \mathcal{X}} \{\theta(x) + \sum_{(i,j) \in E} \gamma_{ij}(x_i, x_j)\}\Big] \quad (8)$$

includes dependencies above and beyond the original tree structure.

Proof. Consider a simple chain with three variables (x_1, x_2, x_3), potential function $\theta(x) = \theta_{12}(x_1, x_2) + \theta_{23}(x_2, x_3)$ and perturbations given by $\gamma(x) = \gamma_{12}(x_1, x_2) + \gamma_{23}(x_2, x_3)$. Let $\Gamma(\hat{x}_\alpha)$ be defined as in Theorem 2 and, similarly, for all subsets $\alpha, \beta \subseteq \{1, \ldots, n\}$, let

$$\Gamma(\hat{x}_\alpha | \hat{x}_\beta) = \Big\{\gamma : \ \hat{x}_\alpha \in \arg\max_{x \in X, x_\beta = \hat{x}_\beta} \{\theta(x) + \gamma(x)\}\Big\}$$

the set of perturbation assignments for which $\hat{x}_\alpha$ is optimal if we plug-in values $\hat{x}_\beta$.

We illustrate that $x_1 \perp\!\!\!\perp x_3 | x_2$ need not hold. To this end, consider probabilities:

$$\mathcal{P}(\hat{x}_i | \hat{x}_2) = P_\gamma\left(\Gamma(\hat{x}_i | \hat{x}_2) | \Gamma(\hat{x}_2)\right), \text{ for } i \in \{1, 3\}$$

Note that the set $\Gamma(\hat{x}_1 | \hat{x}_2)$ is governed by the constraint $\theta_{12}(\hat{x}_1, \hat{x}_2) + \gamma_{12}(\hat{x}_1, \hat{x}_2) \geq \max_{x_1}\{\theta_{12}(x_1, \hat{x}_2) + \gamma_{12}(x_1, \hat{x}_2)\}$ and similarly, $\Gamma(\hat{x}_3 | \hat{x}_2)$ is governed by an analogous constraint on γ_{23}. $\Gamma(\hat{x}_2)$, in contrast, involves inequalities that couple all the perturbation variables together: $\max_{x_1}\{\theta_{12}(x_1, \hat{x}_2) + \gamma_{12}(x_1, \hat{x}_2)\} + \max_{x_3}\{\theta_{23}(\hat{x}_2, x_3) + \gamma_{23}(\hat{x}_2, x_3)\} \geq \max_x\{\theta(x) + \gamma_{12}(x_1, x_2) + \gamma_{23}(x_2, x_3)\}$. Since in general these constraints cannot be decomposed as $(\gamma_{12}, \gamma_{23})$, the set is not a product space.

Consider the following example, where $x_i \in \{0, 1\}$ and $\theta_{12}(1, 1) = 1.9$, $\theta_{12}(0, 0) = 1.2$, $\theta_{12}(0, 1) = 1.1$, $\theta_{12}(1, 0) = 0$ and $\theta_{23}(x_2, x_3) = \theta_{12}(x_3, x_2), \forall x_2, x_3$. For $\hat{x}_2 = 1$, $\Gamma(\hat{x}_2)$ includes the constraint $\max\{1.9 +$

$\gamma_{12}(1,1), 1.1 + \gamma_{12}(0,1)\} + \max\{1.9 + \gamma_{23}(1,1), 1.1 + \gamma_{23}(1,0)\} \geq \max\{1.2+\gamma_{12}(0,0), \gamma_{12}(1,0)\}+\max\{1.2+\gamma_{23}(0,0), \gamma_{23}(0,1)\}$. We argue that there exist i.i.d. perturbation distributions over $(\gamma_{12}, \gamma_{23})$ for which the constraint couples the two variables. In particular, if γ_{12} and γ_{23} are uniform in $\{-1, 1\}$ then for $\gamma_{ij} = (\gamma_{ij}(1,1), \gamma_{ij}(0,1), \gamma_{ij}(0,0), \gamma_{ij}(1,0))$, the configurations $(\gamma_{12}, \gamma_{23}) \in \{((1,1,-1,1),(-1,1,1,1)), ((1,1,-1,1),(1,1,-1,1)), ((-1,1,1,1),(1,1,-1,1))\}$, are in $\Gamma(\hat{x}_2)$, but $((-1,1,1,1),(-1,1,1,1))$ is not, thus it cannot be a product space in this case.

As a result, γ_{12} and γ_{23} become dependent if we condition on $\hat{x}_2$ as the maximizing value. In other words, the indicator functions corresponding to $\Gamma(\hat{x}_1|\hat{x}_2)$ and $\Gamma(\hat{x}_3|\hat{x}_2)$ are also dependent if $\gamma \in \Gamma(\hat{x}_2)$. Whenever x_1 and x_3 depend non-trivially on the corresponding perturbation variables, we conclude that $x_1 \not\perp x_3|x_2$. This is typically the case.

□

The importance of this theorem lies in the modeling power of MAP perturbations. Many important machine learning applications, such as pose estimation, consist of tree structured potential functions whose Gibbs distributions assume conditional independence, e.g., the hands and legs are independent given the body position. Many of these assumptions are made for computation reasons and perturbation models may be able to capture the longer-range dependencies between the parts without increasing the complexity of the potential function.

4 Learning Perturbation Models

So far we have considered tree-structured potential functions and perturbations along the edges in the tree. The resulting induced distributions can be tailored to respect the same Markov structure but, as shown, additional dependencies will emerge in the general case. We will explore here more complex potential functions as well as perturbations that are no longer independent across the edges. The induced distributions in this case will likely involve interactions of all orders. We propose to take advantage of this modeling power and directly learn dependent perturbation models from data.

To specify the perturbation models we use potential functions defined by parameters w and statistics $\phi(x)$, i.e., $\theta(x; w) = w^T \phi(x)$. In contrast to additive perturbations considered earlier, we view w directly as a random variable. The distribution $p(w; \eta)$ governs the randomization and η are the (hyper-)parameters we aim to learn. In the additive case, we would simply use $w = w_0 + \gamma$ where γ is a vector of random perturbations.

The induced distribution over the product space $\mathcal{X}$ is now given by:

$$\mathcal{P}(\hat{x}; \eta) = \int p(w; \eta) [\![\hat{x} = \arg\max_x \theta(x; w)]\!] dw \qquad (9)$$

The goal is to learn the hyper-parameters η that maximize the induced log-likelihood of the data $\sum_{\hat{x} \in \mathcal{S}} \log \mathcal{P}(\hat{x}; \eta)$. This is a latent variable model with continues hidden variables w. In principle, we could use the EM algorithm resulting in the following iterative updates

$$\eta^{(t+1)} = \arg\max_\eta \sum_{\hat{x} \in \mathcal{S}} E_{w \sim p(w|\hat{x}; \eta^{(t)})} \big[\log p(w; \eta) \big] \qquad (10)$$

However, evaluating the expectation requires sampling from the inverse set $\Gamma(\hat{x}) = \{w \mid \hat{x} = \arg\max_x w^T \phi(x)\}$. This is often difficult, requiring costly MCMC methods. We will instead replace the expectation in the E-step with a maximization over w, obtaining a single point in the inverse set $\Gamma(\hat{x})$. This hard-EM algorithm is given by

$$\eta^{(t+1)} = \arg\max_\eta \sum_{\hat{x} \in \mathcal{S}} \max_{w \in \Gamma(\hat{x})} \log p(w; \eta) \qquad (11)$$

The inner maximization remains challenging since the number of constraints specifying the inverse set can be exponential in the number of variables. For example, we might need to enforce $w^\top \phi(\hat{x}) \geq w^\top \phi(x)$ for every $x \in dom(\phi)$. We will show below that there are many problems of interest for which the inverse set can be described compactly.

4.1 Inverse Optimization

Optimization problems over discrete sets such as maximization of $w^\top \phi(x)$ over $x \in dom(\phi)$, can be cast as continuous optimization problems over the corresponding convex hull $conv(\{\phi(x) : x \in dom(\phi)\})$. The convex hull is a polytope defined by linear constraints $\{x : Az \leq b, z \geq 0\}$, and the vertexes of this polytope are exactly the statistics $\phi(x)$. Thus $w \in \Gamma(\hat{z})$ if and only if $\hat{z}$ is the maximizer of the linear objective $f(z) = w^\top z$ over the polytope. In many cases, the constraint matrix A is *totally unimodular*.

Naively one may verify that $\hat{z}$ is the maximizer by trying all the extreme points. More efficiently, we appeal to convex duality in order to maintain a certificate of optimality for $\hat{z}$. A dual certificate is a dual feasible solution that satisfies the complementary slackness constraints: if $\hat{z}_i > 0$ then the corresponding constraint on the dual variable y_i is satisfied with equality $[A^T y]_i = w_i$, and if $[A\hat{z}]_i < b_i$ then $y_i = 0$. Using

the dual certificate, we can maintain the optimality of $\hat{z}$ while changing w. Specifically, we write the inner maximization problem in (11) as a convex program:

$$\max_{w,y} \quad \log p(w;\eta) \tag{12}$$

$$s.t. \quad A^T y \geq w, y \geq 0 \tag{13}$$

$$y_i = 0, \text{ for } i \in \{i|[A\hat{z}]_i < b\} \tag{14}$$

$$[A^T y]_j = w_j, \text{ for } j \in \{j|\hat{z}_j > 0\} \tag{15}$$

Such inverse linear programs have been used before in operations research. The goal is typically to find the parameter setting closest to a given w_0 while ensuring that $\hat{z}$ remains optimal. The distance is a weighted L_p norm, mostly L_1 and L_∞ norms [1]. Also see [4] for a related usage. In our case, $p(w;\eta)$ is a multivariate Gaussian and thus the resulting convex program is quadratic, solved using standard QP solvers.

When the linear program (LP) admits a compact representation, we can represent the inverse set compactly as well since there is a dual variable for every primal constraint. Cases of interest to us include bipartite matching, maximum spanning tree, and so on. When the LP formulation is a relaxation, the constraints (14-15) are tighter than necessary. The inverse program will return a point within a smaller set contained in the inverse set $\Gamma(\hat{x})$ (or the empty set).

We describe below a few examples that are relevant for our models.

Example 1: Image Matching

We start with an assignment problem. For a graph $G = (I \cup J, E, w), E \subseteq I \times J$ with edges weighted by w_{ij} and $|I| = |J| = n$, the goal is to find the maximum weight matching that assigns each element in I to exactly one element in J. Document ranking and key-point matching in images can be modeled as assignment problems.

By reweighing the edges, the optimal assignment can be formulated as a minimum cost matching problem, which can be computed in polynomial time using the Hungarian algorithm [31]. Note that sampling and computing the partition function remain #P-complete [36] though MCMC-based fully-polynomial approximation schemes exist [18]. In comparison, perturbation models rely only on the efficient polynomial time maximization.

The minimum cost matching can be obtained by minimizing a linear objective $f(z) = w^T z$ subject to constraints. The constraints ensure that each vertex is incident to exactly one edge in the matching $\sum_{k\in I} z_{kj} = 1, \sum_{k\in J} z_{ik} = 1$ [31]. Using dual certificates, we can formulate the inverse problem, i.e., $\max_{w\in\Gamma(\hat{z})} \log p(w;\eta)$ as a convex program:

$$\max_{w,u,v} \quad \log p(w;\eta)$$

$$s.t. \quad u_i + v_j = w_{ij}, \ (i,j) \in \{(i,j)|\hat{z}_{ij} \neq 0\}$$

$$u_i + v_j \leq w_{ij}, \ (i,j) \in \{(i,j)|\hat{z}_{ij} = 0\}$$

where $\hat{z}$ is the observed assignment and u and v are dual variables. The compact description involves $2n^2$ constraints and $2n$ additional (dual) variables.

Example 2: Pose Estimation

In pose estimation, the human body is modeled as a tree-structured graphical model, where nodes correspond to body parts. The highest scoring labeling specifies the estimated locations for the parts [41]. The tree structure is computationally appealing, but it assumes that limbs are independent given the body position. Perturbation models can capture longer range dependencies even when the potential function corresponds to a tree.

While inference and sampling in tree-structured models is easy, sampling from the inverse set is difficult. The constraints enforcing the solution $\hat{x}$ to be optimal extend beyond the tree structure. The MAP solution can be nevertheless cast as a maximization of a linear objective $f(\mu) = w^T\mu$ over the local polytope $\mathcal{M}_L(G) = \{\mu \geq 0 | \sum_{x_j} \mu_{i,j;x_i,x_j} = \mu_{i;x_i} \forall i,j,x_i, \sum_{x_i} \mu_{i,j;x_i,x_j} = \mu_{j;x_j} \ \forall i,j,x_j, \ \sum_{x_i} \mu_{i;x_i} = 1 \ \forall i\}$. For trees, the solution $\hat{\mu}$ is integral and corresponds to the maximum assignment $\hat{x}$ [6]. In other words, $\hat{\mu}$ describes $\hat{x}$ in terms of local marginals. Using dual certificates, we can write the inverse problem as:

$$\max_{w} \quad \log p(w;\eta)$$

$$s.t. \quad y_i - \sum_j y'_{i,j;x_i} - \sum_j y''_{j,i;x_i} \geq w_{i;x_i}, \text{for } \hat{\mu}_{i;x_i} = 0$$

$$y_i - \sum_j y'_{i,j;x_i} - \sum_j y''_{j,i;x_i} = w_{i;x_i}, \text{for } \hat{\mu}_{i;x_i} > 0$$

$$y'_{i,j;x_i} + y''_{i,j;x_j} \geq w_{i,j;x_i,x_j}, \text{for } \hat{\mu}_{i,j;x_i,x_j} = 0$$

$$y'_{i,j;x_i} + y''_{i,j;x_j} = w_{i,j;x_i,x_j}, \text{for } \hat{\mu}_{i,j;x_i,x_j} > 0$$

where y, y', y'' are dual variables corresponding to the marginal constraints. The constraints are satisfied with equality when the corresponding marginals in $\hat{\mu}$ are non-zero.

Example 3: Image Segmentation

Image segmentation and other computer vision tasks can be modeled as energy minimization problems with sub-modular potentials. Minimum graph cuts are used as tools for finding the optimal assignments [33].

For a graph $G = (V, E, w)$ with $|V| = n, |E| = m$ and edge costs given by w, the minimum s-t cut problem aims to find a subset of vertices S, with $s \in S$ and $t \in V \setminus S$, such that the cost of the cut (weight of the edges crossing S and $V \setminus S$) is minimized. The dual problem is maximum-flow, and we can solve the inverse problem via

$$\begin{aligned} \max_{w,y} \quad & \log p(w; \eta) \\ s.t. \quad & \sum_i y_{ik} = \sum_j y_{kj}, \ \forall k \neq s, k \neq t \\ & 0 \leq y_{ij} \leq w_{ij}, \forall i, j \\ & y_{ij} = w_{ij}, \text{ for } (i,j) \in \{(i,j) | \hat{z}_{ij} > 0\} \end{aligned}$$

where y are the dual variables and $\hat{z}$ encodes the observed cut. We obtain a compact, polynomial size representation of the inverse problem, at the cost of introducing m additional variables. For image segmentation and for most examples we provide, the number of additional variables is at most the number of parameters w.

Example 4: Natural Language Parsing

Dependency parsing can be formulated as a maximum directed spanning tree problem over the words in the sentence [25]. Different interpretations of the sentence correspond to different parse trees. As a result, the target parse can be inherently ambiguous. Perturbation models can be used to efficiently sample high-scoring parse trees to represent candidate interpretations.

In this case, a polynomial size representation of the inverse problem can be obtained via LP formulation of the minimum cost directed tree problem. In a graph $G = (V, E, w)$, the primal LP involves minimizing a linear objective $\sum_{(i,j)\in E} w_{ij} z_{ij}$ subject to constraints ensuring that for every node $u \in V \setminus \{r\}$ there is an r–u flow $f^{(u)}$ of value 1 with $f^{(u)}_{ij} \leq z_{ij}$ [31]. The feasible set is the projection of a high dimensional polytope in mn dimensions, governed by at most $n(2m + n)$ constraints. Here n and m are the length of the sentence and the number of edges, respectively. As a result, using the dual certificate approach (omitted), we can formulate the inverse problem with $O(mn)$ additional variables.

4.2 Penalty-based Inverse Optimization

The inverse optimization framework provides a clean way of solving the inner maximization in (11) for many problems of interest. For completeness, we also provide examples where the size of the LP formulation is large relative to the number of parameters in w.

Consider learning a perturbation model over binary images of size $k \times k$, guided by a potential function $\theta(x; w) = \sum_{i=1}^{n} w_i x_i + \sum_{(i,j)\in E} w_{ij} x_i x_j$, $|E| = m$. For large k, it may be impractical to learn both unary and pairwise potentials resulting in $n + m$ parameters. We can instead estimate a subset of parameters, e.g. fix the higher-order potentials and learn n parameters for node potentials. Nonetheless, the min-cut inverse LP formulation in Example 3 adds additional variables for each edge and even for estimating a subset of parameters, the number of variables is given by $n + m$.

In many cases we must resort to constraints of the form $w^T \phi(\hat{x}) \geq w^T \phi(x), \forall x$. Assuming that the perturbations follow a multivariate Gaussian distribution, the inverse optimization problem is quadratic

$$\min_w \ (w - \mu)^T \Sigma^{-1} (w - \mu) + C \left[\max_x w^T \phi(x) - w^T \phi(\hat{x}) \right]$$

The objective is similar to structured SVM [35] and a similar approach has been explored in [34]. The problem can be solved using typical methods for structured SVMs, such as cutting-planes or gradient descent methods. We illustrate this in the experimental section using a sub-gradient descent with a decreasing step size.

5 Experiments

The goal of our experiments is to demonstrate that perturbation models capture dependencies above and beyond the original structure of the potential function and to illustrate the duality approach for learning. We first exemplify the induced dependencies on a simple image modeling task and then we apply the hard-EM framework in the context of image matching.

5.1 Image segmentation

We selected four images from the Large Binary Image Database[1] representing basketball player silhouettes, with the goal of learning a model over the basketball player poses and showing that perturbation models are able to store multiple modes and sample from them.

We used an Ising model over labels $y_i \in \{+1, -1\}$ with potentials $\theta(y_i)$ encoding whether pixel i is foreground or background and $\theta(y_i, y_j)$ encouraging adjacent pixels to have the same labels. We assumed $\theta(y_i, y_j) = y_i y_j, \theta_i(y_i) = \gamma_i y_i$ and learned a distribution over the node parameters γ_i. Since the model contained node potentials only (resulting in 2500 parameters), we solved the inverse problem using the sub-gradient approach explained in the previous section. For each iteration of the hard-EM algorithm, we performed 3 iterations of the sub-gradient algorithm for

[1] `http://www.lems.brown.edu/~dmc/`

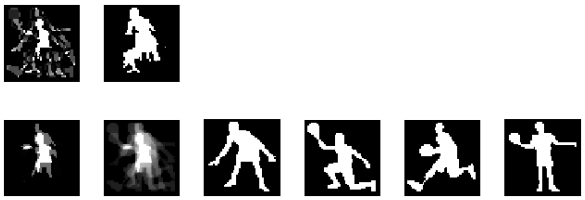

Figure 1: First line: max-margin parameters and resulting segmentation, second line: the mean of the perturbation parameters, the average segmentation and the four images with the highest count.

each example, initialized with the point estimate from the previous hard-EM iteration. Since the setting is so simple, the hard-EM algorithm converged in less than 20 iterations. For computing the maximum likelihood estimates of η in the M-step we performed regularization by adding a constant c to the diagonal elements of the estimated covariance matrix (we set c to 0.1). We also implemented a structural SVM approach, using a similar stochastic sub-gradient algorithm.

In Figure 1, second line, we show in this order the mean of the perturbation parameters γ, the average segmentation from 10^4 samples and the four images with the highest count. In this case, the four images correspond to the four human poses we considered and images visually similar to them obtain a similar score. The first line shows the learned node parameters and the max-margin maximum weight configuration.

The potential function encodes only local interactions through the lattice structure, but the induced distribution shows longer range dependencies. This is due to the correlations in the latent space as illustrated in Figure 2. For pixels that are always foreground or background the covariance matrix reveals no correlations. The others have strong positive correlations with pixels that are only activated on the same pose, and negative correlations with other poses. To further understand the perturbation models we look at independent samples, Figure 3, where the perturbation distribution is a multivariate gaussian with unrestricted, resp. diagonal, covariance matrix (first two lines). The second model captures few or no long-range dependencies in this case.

Instead of perturbation models, one may learn a multivariate gaussian model over the binary images and compute a sample image by thresholding each pixel independently. We also show samples from these models in Figure 3, last two line, where the covariance matrix is unrestricted, resp. diagonal. The latent space is capturing the long-range correlations, but the lack of structure in the MAP solver results in visual artifacts.

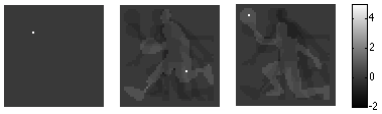

Figure 2: Correlations between a reference pixel (white) and the rest, as captured by the covariance matrix of the perturbation distribution. We show a pixel that is always off (so no correlations) and two pixels that are activated on different poses.

5.2 Image matching

We illustrate the LP duality approach for a matching task on images from the Buffy Stickmen dataset[2]. Each frame is annotated with segment locations for six body parts and we use the framework of [41] to enlarge this set of locations such that we obtain 18 keypoints per image. We select frames of the same person throughout an episode and from the resulting set of all image pairs we randomly select two disjoint sets for training and testing (15 train pairs and 23 test pairs). The set of keypoints for an image pair serves as the ground truth for our matching experiments.

We represent the matching as a permutation of keypoints denoted by π, and assume the following potential function, following [37], $\theta(I, I', \pi; w) = \sum_{i,j} w^T(\psi(I, i) - \psi(I', j))^2$. The features $\psi(I, k)$ are the SIFT descriptors evaluated keypoint k.

The inference problem can be formulated as an assignment problem, so we learn the perturbation distribution using the hard-EM algorithm, and computing the point estimate using the inverse optimization formulation. In this case, the inverse problem becomes a quadratic program with 26 additional variables and 324 constraints corresponding to edges.

Figure 4 shows an example pair from the test set. We extract SIFT features at scale 5 and we return the matching with the highest count after 1000 samples. In this case the perturbation model shows similar performance with SVM: the average error of the perturbation model after 1000 samples was equal to 8.47 while the average error of max margin was 8.69.

6 Related Work

The Gibbs distribution plays a key role in many areas of computer science, statistics and physics. To learn more about its roles in machine learning we refer the interested reader to [19, 38]. The Gibbs distribution as well as its Markov properties can be realized from the statistics of high dimensional random MAP perturba-

[2] http://www.robots.ox.ac.uk/~vgg/data/stickmen/

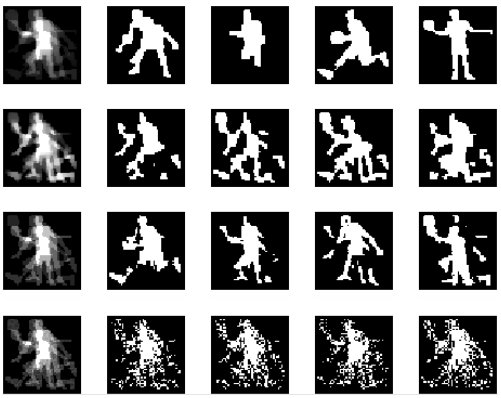

Figure 3: The average segmentation and samples from four models, one per line: perturbation model where the perturbations have unrestricted vs. diagonal covariance matrix and multivariate gaussian model with unrestricted vs. diagonal covariance matrix.

tions with the Gumbel distribution (see Theorem 1), [29, 34, 13, 14]. For comprehensive introduction to extreme value statistics we refer the reader to [22].

Recent work [28, 29, 34] explores the different aspects of low dimensional MAP perturbation models. Papandreou et al. [28] describe sampling from the Gaussian distribution with random Gaussian perturbations. Later [29], they show empirically that MAP predictors with low dimensional perturbations share similar statistics as the Gibbs distribution. In our work we investigate the dependencies of such probability models. Specifically, we present non-i.i.d. low dimensional random perturbations that recover the Markov properties of tree structured Markov random fields. We also show that independent low dimensional perturbations may model long-range interactions. Tarlow et al. [34] describe the Bayesian perspectives of these models and their efficient sampling procedures, as well as several learning techniques including hard-EM. In contrast, we focus on understanding the structure of the induced distribution and our learning approach is different. We use dual LPs in our hard-EM approach so as to obtain compact representations of the inverse polytope when possible, while Tarlow et. al [33] focus on cutting plane approaches. When using cutting plane approaches for only a couple of iterations, the hard-EM estimates often fall outside the inverse polytope. Our dual LP approach alleviates this problem and in our experiments almost all estimates fall within the inverse polytope.

Our experiments show that we are able to sample from the modes of the distribution. Alternatively, one may use the M-best approach and its diverse-versions to recover such modes [42, 7, 2, 26, 3, 11]. Finding the

Figure 4: Example matching returned by the randomized MAP model. This is the matching with the highest count from 100 samples and has error equal to 4.

M-best carries a computational effort which extends beyond our learning approach whose complexity is as a 1-best solver. Alternatively, one may sample from determinantal point processes to retrieve the modes of the distributions [23]. This learning approach concerns problems that can be described by determinants while our approach is based on MRF potentials.

7 Discussion

This work explored random MAP perturbation models. We showed that perturbation models can be tailored to represent tree structured models but also that they typically would involve long-range dependencies above and beyond the original structure. Perturbation models can be viewed as latent variable models and we demonstrated distributions over perturbations can be learned using a hard-EM approach. In the E-step, an inverse convex program is used to confine the randomization to the parameter polytope responsible for generating the observed assignment. We illustrated the approach on several computer vision problems.

This work can be extended in many ways. A complete understanding of conditioning in perturbation models is missing, as is a full account of long-range dependences. Unlike sampling, evaluating the MAP assignment from an induced model is not straightforward. Finally, models with dependent perturbations, while seemingly powerful, are not yet well-understood.

Acknowledgements

The work was supported in part by Google Inc (Rethinking AI) and Skolkovo foundation (Machine learning for BigData).

References

[1] Ravindra K Ahuja and James B Orlin. Inverse optimization. In *Operations Research*, 2001.

[2] D. Batra. An efficient message-passing algorithm for the m-best map problem. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[3] Dhruv Batra, Payman Yadollahpour, Abner Guzman-Rivera, and Gregory Shakhnarovich. Diverse m-best solutions in markov random fields. In *ECCV*, 2012.

[4] Vassil Chatalbashev. Inverse convex optimization.

[5] P.F. Felzenszwalb and R. Zabih. Dynamic programming and graph algorithms in computer vision. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):721–740, 2011.

[6] Menachem Fromer and Amir Globerson. An lp view of the m-best map problem. *Advances in Neural Information Processing Systems (NIPS)*, 2009.

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

[8] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1984.

[9] L.A. Goldberg and M. Jerrum. The complexity of ferromagnetic ising with local fields. *Combinatorics Probability and Computing*, 16(1):43, 2007.

[10] E.J. Gumbel and J. Lieblein. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Govt. Print. Office, 1954.

[11] Abner Guzman-Rivera, Pushmeet Kohli, and Dhruv Batra. Faster training of structural svms with diverse m-best cutting-planes. In *Discrete Optimization in Machine Learning Workshop (DISCML-NIPS)*, 2012.

[12] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[13] T. Hazan and T. Jaakkola. On the partition function and random maximum a-posteriori perturbations. *ICML*, 2012.

[14] T. Hazan, S. Maji, and T. Jaakkola. On sampling from the gibbs distribution with random maximum a-posteriori perturbations. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[15] Mark Huber. A bounding chain for swendsen-wang. *Random Structures and Algorithms*, 2003.

[16] M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

[17] M. Jerrum, A. Sinclair, and E. Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.

[18] Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM (JACM)*, 51(4):671–697, 2004.

[19] D. Koller and N. Friedman. *Probabilistic graphical models*. MIT press, 2009.

[20] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 2006.

[21] T. Koo, A.M. Rush, M. Collins, T. Jaakkola, and D. Sontag. Dual decomposition for parsing with non-projective head automata. In *EMNLP*, 2010.

[22] S. Kotz and S. Nadarajah. *Extreme value distributions: theory and applications*. World Scientific Publishing Company, 2000.

[23] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 2012.

[24] S. Maji, T. Hazan, and T. Jaakkola. Efficient boundary annotation using random map perturbations. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014.

[25] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. Non-projective dependency parsing using spanning tree algorithms. In *EMNLP*, 2005.

[26] E. Mezuman, D. Tarlow, A. Globerson, and Y. Weiss. Tighter linear program relaxations for high order graphical models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

[27] F. Orabona, T. Hazan, A Sarwate, and T. Jaakkola. On measure concentration of random maximum a-posteriori perturbations. In *ICML*, 2014.

[28] G. Papandreou and A. Yuille. Gaussian sampling by local perturbations. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[29] G. Papandreou and A. Yuille. Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *ICCV*, 2011.

[30] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, 1988.

[31] Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*. Springer, 2003.

[32] D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

[33] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1068–1080, 2007.

[34] D. Tarlow, R.P. Adams, and R.S. Zemel. Randomized optimum models for structured prediction. In *AISTATS*, 2012.

[35] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, page 104. ACM, 2004.

[36] L.G. Valiant. The complexity of computing the permanent. *Theoretical computer science*, 1979.

[37] Maksims Volkovs and Richard S Zemel. Efficient sampling for bipartite matching problems. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[38] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008.

[39] JianSheng Wang and RH Swendsen. Nonuniversal critical dynamics in monte carlo simulations. *Physical review letters*, 1987.

[40] Tomás Werner. High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimization (map-mrf). In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[41] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[42] Chen Yanover and Yair Weiss. Finding the m most probable configurations using loopy belief propagation. *Advances in Neural Information Processing Systems (NIPS)*, 2004.