Advanced Structured Prediction

Editors:

Tamir Hazan tamir.hazan@technion.ac.il
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel

George Papandreou gpapan@google.com
Google Inc.
340 Main St., Los Angeles, CA 90291 USA

Daniel Tarlow dtarlow@microsoft.com
Microsoft Research
Cambridge, CB1 2FB, United Kingdom

The MIT Press
Cambridge, Massachusetts
London, England

1 Learning with Maximum A-Posteriori Perturbation Models

Andreea Gane agane@csail.mit.edu
CSAIL, MIT
Cambridge, MA, USA

Tamir Hazan tamir.hazan@technion.ac.il
Technion - Israel Institute of Technology
Haifa, Israel

Tommi Jaakkola tommi@csail.mit.edu
CSAIL, MIT
Cambridge, MA, USA

Perturbation models are families of distributions induced from perturbations. They combine randomization of the parameters with maximization to draw unbiased samples. In this chapter, we describe randomization both as a modeling tool and as a means to enforce diversity and robustness in parameter learning. A perturbation model defined on the basis of low order statistics typically introduces high order dependencies in the samples. We analyze these dependencies and seek to estimate them from data. In doing so, we shift the modeling focus from the parameters of the potential function (base model) to the space of perturbations. We show how to estimate dependent perturbations over the parameters using a hard EM approach, cast in the form of inverse convex programs and illustrate the method on several computer vision problems.

1.1 Introduction

In applications that involve structured objects, such as object boundaries, textual descriptions, or speech utterances, the key problem is finding expressive yet tractable models. In these cases, the likely assignments are guided by potential functions over subsets of variables. The feasibility of inference is typically linked to the structure of the potential function and the tradeoff is between rich, faithful models defined on complex potential functions on one hand, and limited but manageable models on the other.

For instance, in natural language parsing, the goal is to return a dependency tree where arcs encode dependency relations, such as between a predicate and its subject. Whenever the interactions are of high order, computing the dependency tree corresponds to an NP-hard combinatorial optimization problem (McDonald and Satta, 2007), but when resorting to tractable formulations by limiting the type of interactions, the expressive power of the model is limited. In general, most realistic models for natural language parsing (Koo et al., 2010a), speech recognition (Rabiner and Juang, 1993) or image segmentation/captioning (Nowozin and Lampert, 2011; Fang et al., 2015) involve interactions between distant words in the sequence or large pixel neighborhoods.

Typical probabilistic models defined on structured potential functions make use of the Gibbs' distribution and its properties. Specifically, the structure of the potential function can be encoded as a graph that specifies conditional independencies (Markov properties) among the variables: two sets of vertices in the graph are conditionally independent when they are separated by observed vertices (e.g., Wainwright and Jordan (2008); Koller and Friedman (2009)). These assumptions are central for designing efficient exact or approximate inference techniques. Successful methods exploiting them include belief propagation (Pearl, 1988), Gibbs sampling (Geman and Geman, 1984), Metropolis-Hastings (Hastings, 1970) or Swendsen-Wang (Wang and Swendsen, 1987). In specific cases one can sample efficiently from a Markov random field model by constructing a rapidly mixing Markov chain (cf. (Jerrum and Sinclair, 1993; Jerrum et al., 2004a; Huber, 2003)). 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 (Jerrum and Sinclair, 1993; Goldberg and Jerrum, 2007).

Finding a single most likely assignment (MAP) structure is 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 (Kolmogorov, 2006) or by devising linear programming relaxations and successively refining them (Sontag et al., 2008; Werner, 2008). Furthermore, even when computing the MAP is provably hard, approximate techniques, such as loopy belief propagation (Murphy et al., 1999), tree reweighed message passing (Wainwright et al., 2005), local search algorithms (Zhang et al., 2014) or convex relaxations (Koo et al., 2010b) are often successful in recovering the optimal solutions (Koo et al., 2010a).

Recently, MAP inference has been combined with randomization to define new classes of probability models that are easy to sample from (Papandreou and Yuille, 2011; Tarlow et al., 2012; Hazan and Jaakkola, 2012; Hazan et al., 2013; Orabona et al., 2014; Maji et al., 2014). Each sample from these perturbation models involve randomization of Gibbs' potentials and finding the corresponding maximizing assignment. The models are shown to provide unbiased samples from the Gibbs distribution when perturbations are independent across assignments (Papandreou and Yuille, 2011; Tarlow et al., 2012) and have been applied to several applications where the underlying combinatorial problem is easy to optimize, but difficult to sample from, including boundary annotation (Maji et al., 2014) and image partitioning (Kappes et al., 2015), but having a full account of the properties and power of perturbation models remains an open problem.

In this chapter, we describe and extend our work (Gane et al., 2014) on understanding and exploiting the expressive power of perturbation models. Specifically, the properties of induced distribution are heavily governed by randomization. In contrast to Gibbs' distributions, low order potentials, after undergoing randomization and maximization, lead to high order dependencies in the induced distributions. Furthermore, while conditioning in Gibbs' distributions is straightforward, conditioning in perturbation models implies restricting the randomizations to a non-trivial set and performing this efficiently is still an open problem.

Finally, we explore the interplay between learning algorithms and tractability of inference procedures on complex potential functions. We describe dependent perturbations as a modeling tool for learning models which lead to tractable inference. 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.

1.2 Background and Notation

In this chapter we are concerned with modeling distributions over structured objects $x \in \mathcal{X}$, such as image segmentations and keypoint matchings, where $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ is a discrete product space. We are scoring the possible assignments via a real valued potential function $\theta(x) = \theta(x_1, ..., x_n)$, where excluded configurations are implicitly encoded by setting $\theta(x) = -\infty$ whenever $x \notin dom(\theta)$. For instance, a foreground-background segmentation over an image of size $n \times m$ can be encoded by $x = (x_{ij})_{i\in[n],j\in[m]} \in \{0,1\}^{n\times m}$, where $x_{ij} = 1$ denotes a foreground pixel at position (i, j). If we want to explicitly encode that a foreground object is always present, then $\theta(x) = -\infty$ whenever $x_{ij} = 0, \forall i \in [n], j \in [m]$.

Since dealing with arbitrary scoring functions is computationally intractable, $\theta(x)$ is typically defined as a sum of local potentials $\theta(x) = \sum_{\alpha\in\mathcal{A}} \theta_\alpha(x)$, where α denotes a small subset of variables (factor) and $\mathcal{A}$ denotes the set of all such factors. In the image segmentation case, $\mathcal{A}$ may include local neighborhoods of the form $\{(i+d_i, j+d_j)|d_i, d_j \in \{+1, 0, -1\}\}$. In the following, we will often skip specifying $\mathcal{A}$ and write $\theta(x) = \sum_\alpha \theta_\alpha(x)$ for simplicity.

Traditionally, the potentials are mapped to the probability scale via the Gibbs' distribution:

$$p(x_1, ..., x_n) = \frac{1}{Z(\theta)} \exp(\theta(x_1, ..., x_n)) \tag{1.1}$$

Distributions defined in this manner have a number of desirable properties. For instance, the maximum-a-posteriory (MAP) prediction corresponds to the highest scoring assignment $\hat{x} = \arg\max_x \theta(x)$, the set of conditional dependencies can be read from the structure of the potential function, and the model can be easily extended to handle partially observed data. Unfortunately, such distributions are challenging to learn and sample from, depending on how the potential function decomposes.

Our approach is based on randomizing potentials in Gibbs' distributions. We add a random function $\gamma : \mathcal{X} \to R$ to the potential function and draw samples by solving the resulting MAP prediction problem:

$$x^* = \arg\max_{x\in\mathcal{X}} \{\theta(x) + \gamma(x)\}. \tag{1.2}$$

The distribution induced by the samples is given by

$$\mathcal{P}(\hat{x}) = P_\gamma\Big[\hat{x} \in \arg\max_{x\in X} \{\theta(x) + \gamma(x)\}\Big] \tag{1.3}$$

and its properties are heavily dependent on the nature of randomization.

The simplest approach to designing the perturbation function γ 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. Specifically, due to the max-stability property of the Gumbel distribution, one can preserve the Markov properties of the Gibbs model. However, each realization x^* in this setup requires an independent draw of $\gamma(x)$, $x \in \mathcal{X}$, i.e., a high dimensional randomization.

Theorem 1.1. *(Gumbel and Lieblein, 1954) 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*

$$P_\gamma\Big[\hat{x} \in \arg\max_{x\in\mathcal{X}}\{\theta(x) + \gamma(x)\}\Big] = \frac{1}{Z(\theta)}\exp(\theta(\hat{x})) \tag{1.4}$$

However, this construction requires for every sample the instantiation of $|\mathcal{X}|$ random variables $(\gamma(x))_{x\in\mathcal{X}}$, which is not feasible in practice. Since perturbation models are useful only if they can be succinctly parametrized, our focus is on investigating *low-dimensional perturbations* which have the same structure as the potential function:

$$\mathcal{P}(\hat{x}) = P_\gamma\Big[\hat{x} \in \arg\max_{x\in X}\big\{\sum_\alpha(\theta_\alpha(x_\alpha) + \gamma_\alpha(x_\alpha))\big\}\Big] \tag{1.5}$$

In this case, each sample requires instantiating $\gamma_\alpha(x_\alpha)$ for each α and each assignment x_α, which is typically a much smaller set. Finally, since the noise function shares the structure of the potential function, the optimization algorithms designed for the original potential function remain applicable. We will often refer to the new (randomized) potential function as $\tilde{\theta}(x) = \sum_\alpha \tilde{\theta}_\alpha(x_\alpha)$, where $\tilde{\theta}_\alpha(x_\alpha) = \theta_\alpha(x_\alpha) + \gamma_\alpha(x_\alpha)$.

1.3 Expressive Power of Perturbation Models

Perturbation models were originally introduced as a way to approximate intractable Gibbs' distributions. In this chapter, we use perturbation models as a modeling tool, seeking to understand their properties, and how to estimate them from data.

The idea of specifying distributions over combinatorial objects by linking randomization and combinatorial optimization is not inherently limiting. At one extreme, the randomization may correspond to samples from the target distribution itself. Of course, the combination is advantageous only when both the randomization and the associated combinatorial problem are

tractable. To this end, we focus on randomizing potentials in Gibbs' distributions whose MAP assignment can be obtained in polynomial time. The randomization we introduce will therefore have to respect how the potential functions decompose. For example, randomization of $\theta(x) = \sum_\alpha \theta_\alpha(x)$ should only directly affect individual terms $\theta_\alpha(x)$.

One of the key questions we address is how the resulting perturbation models differ from the associated Gibbs' models that they are based on. Gibbs' distributions are naturally understood in terms of Markov properties. Will these carry over to perturbation models as well? We will show that in contrast to Gibbs' distributions, low order potentials, after undergoing randomization and maximization, lead to high order dependencies in the induced distributions. Such induced dependences can be viewed as additional modeling power and specifically exploited and learned from data. Markov properties can be enforced in special cases such as with tailored perturbations in tree structured models, if desired.

Perturbation models yield simple mechanisms for drawing unbiased samples but they are cumbersome with respect to conditioning. Indeed, "plug-in" conditioning natural in Gibbs' distributions does not carry over to perturbation models. Conditioning requires care, restricting the randomization such that the setting of the observed variables are indeed obtained as part of maximizing assignments. We show how this can be done in simple examples.

1.4 Higher Order Dependencies

In this section, we show that perturbation models defined via low dimensional randomizations do not follow the Markov-type dependencies inherent in Gibbs distributions. We focus on perturbation models with tree structured potential functions and edge-based randomization, 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 1.2. *Most perturbation models with tree structured potential functions and i.i.d. perturbation variables* $\{\gamma_{ij}(x_i, x_j)\}$ *indexed by* $(i,j) \in E, (x_i, x_j) \in \mathcal{X}_i \times \mathcal{X}_j$ *result in an induced model (1.5) that 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

$$\Gamma(\hat{x}_\alpha) = \left\{\gamma : \ \hat{x}_\alpha \in \arg\max_{x \in X}\{\theta(x) + \gamma(x)\}\right\} \tag{1.6}$$

and, similarly, for all subsets $\alpha, \beta \subseteq \{1, \ldots, n\}$, let

$$\Gamma(\hat{x}_\alpha|\hat{x}_\beta) = \left\{\gamma : \ \hat{x}_\alpha \in \arg\max_{x \in X, x_\beta = \hat{x}_\beta} \{\theta(x) + \gamma(x)\}\right\} \tag{1.7}$$

be 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 $\theta_{23}(a, b) = \theta_{12}(b, a), \forall a, b \in \{0, 1\}$. 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 $\gamma_{12}(x_1, x_2) \sim U\{-1, 1\}\ \forall (x_1, x_2) \in \{0, 1\}^2$, $\gamma_{23}(x_2, x_3) \sim U\{-1, 1\}\ \forall (x_2, x_3) \in \{0, 1\}^2$ and U is the uniform distribution, 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 $x_1 \not\perp\!\!\!\perp x_3|x_2$. This is typically the case.

□

The key role of this theorem is to highlight how perturbation models might posses higher modeling power than their Gibbs counterparts. The choice of tree structured potential functions is often guided by computational reasons,

rather than the need for conditional independence. Specifically, in a pose estimation application the goal is to relate a set of keypoints $x = (x_i)_{i\in[n]}$, where dimensions x_i are (pixel) locations arms, legs, body trunk or head and n is the total number of keypoints. A typical scoring function is $\theta(x) = \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j)$, where the set of edges E includes pairs such as pair-trunk, arm-trunk, leg-trunk and the local scores $\theta_{ij}(x_i, x_j)$ depend on the distance between the keypoints. From the structure of $\theta(x)$, the Gibbs' distribution implies that the limbs locations are independent given the trunk. Perturbation models have the potential to capture additional long range dependencies between the parts without increasing the complexity of the scoring function.

1.5 Markov Properties and Perturbation Models

Given that typically low order perturbations lead to high order dependencies, we ask whether enforcing the Markov properties is possible in this case.

In the simplest case, whenever the Gibbs distribution is independent, it can indeed be represented using low order potentials. Specifically, 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. To show that the perturbation model matches the Gibbs distribution in this case we apply Theorem 1.1 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 show that the tree structured potentials can also be randomized such that the induced distribution corresponds to the Gibbs' distribution.

1.5.1 Tree-Structured Perturbation Models

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:

$$\theta_i(x_i) = \log p(x_i), \quad \theta_{ij}(x_i, x_j) = \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \tag{1.8}$$

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

Theorem 1.3. *Consider the Gibbs distribution with a tree structured Markov random field. Then for any potential function*

$$\theta(x) = \sum_{i=1}^{n} \theta_i(x_i) + \sum_{(i,j)\in E} \theta_{ij}(x_i, x_j) \tag{1.9}$$

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*

$$p(\hat{x}) = P_\gamma\Big[\hat{x} \in \operatorname*{argmax}_{x\in\mathcal{X}} \{\theta(x) + \sum_{(i,j)\in E} \gamma_{ij}(x_i, x_j)\}\Big] \tag{1.10}$$

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 $\gamma_{ij}(x_i, x_j) =$

$$\hat{\gamma}_{ij}(x_i, x_j) - \max_{x'_j} \big\{\theta_{ij}(x_i, x'_j) + \theta_j(x'_j) + \hat{\gamma}_{ij}(x_i, x'_j)\big\} \tag{1.11}$$

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

$$p(\hat{x}_1, \hat{x}_2) = P_\gamma\big[\hat{x}_1, \hat{x}_2 \in \operatorname*{argmax}_{x\in\mathcal{X}} \{\theta(x) + \sum_{(i,j)\in\vec{E}} \gamma_{ij}(x_i, x_j)\}\big] \tag{1.12}$$

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

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

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 1.1 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 $\Gamma(\hat{x}_\alpha)$

the set of perturbation assignments for which $\hat{x}_\alpha$ is optimal, as in (1.7). Recall the vertices are topologically ordered, thus we aim at showing that

$$p(x_j|x_i) = P_\gamma\Big(\Gamma(x_1,...,x_j)|\Gamma(x_1,...,x_{j-1})\Big) \tag{1.13}$$

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 1.1. □

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.

1.6 Conditional Distributions

Modeling and efficiently using conditional distributions are key issues in applications involving partially observed data. These include finding dense correspondences across images when only partial human annotations are provided, combining information from multiple predictors (semi-supervised learning) and so on. In Gibbs' models, regardless of the difficulty of inference calculations, conditioning is typically a straightforward operation, performed by plugging in the observed data. On the other hand, conditioning in perturbation models is a challenging open problem. In this case we cannot merely set the observed variables to their values. Instead, we must ensure that the observed values are selected via global maximization.

Specifically, for any subset of variables $x_\alpha, x_\beta, \alpha \cap \beta = \emptyset, \alpha, \beta \in V$, the conditional $\mathcal{P}(\hat{x}_\alpha|\hat{x}_\beta)$ is obtained by first sampling noise realizations that are consistent with observed data and maximizing the perturbed potential over the remaining variables:

$$\gamma \sim p(\gamma|\gamma \in \Gamma(\hat{x}_\beta)) \tag{1.14}$$

$$\hat{x}_\alpha \leftarrow \operatorname*{argmax}_{x_\alpha} \max_{x_{V\setminus\alpha}} \tilde{\theta}(x) \tag{1.15}$$

Recall that $\Gamma(\hat{x}_\beta)$ is the set of perturbations for which the maximizing argument agrees with $\hat{x}_\beta$. The resulting distribution of $\hat{x}_\alpha$ is typically different from the one obtained by fixing the observed values $\hat{x}_\beta$ while maximizing over the remaining ones:

$$q(\hat{x}_\alpha|\hat{x}_\beta) = \Pr(\hat{x}_\alpha \in \operatorname*{argmax}_{x_\alpha} \max_{x_{V\setminus\{\alpha,\beta\}}} \tilde{\theta}(x_{V\setminus\beta}, \hat{x}_\beta)) \tag{1.16}$$

To show how these two approaches may lead to different induced distributions, consider the example provided in the proof of Theorem 1.2. When conditioning on $\hat{x}_2$ in the three-variable chain $x_1 - x_2 - x_3$, the perturbation variables γ_{12} and γ_{23} become coupled and this is shown to imply conditional dependency between x_1 and x_3. However, in the distribution obtained by fixing the value of $\hat{x}_2$ and sampling γ_{12}, γ_{23} from their original (independent) distributions, x_1 and x_3 become independent. Therefore, the two distributions are in general not the same and in particular, the ability to perform conditioning by "plugging in" the observed variables is related to the higher order dependencies that arise with perturbation models.

In practice, the key difficulty for conditioning in perturbation models stems from dealing with the set $\Gamma(\hat{x}_\beta)$, which is often a union of (disparate) cones. This makes the posterior distribution $p(\gamma|\gamma \in \Gamma(\hat{x}_\beta))$ difficult to describe and sample from.

In the rest of this section, our aim is to characterize models for which we can perform conditioning with respect to a restricted subset of variables. We start by describing model constraints which ensure conditional independence (with respect to a single variable) in a three-variable chain. Furthermore, the conditions can be extended to enforce conditional independence in models whose potential functions decompose along the edges of the tree. We then show that when such conditions are met, we can perform conditioning on a single variable by fixing the observed variable to its value. While the conditions are restrictive, we show that there exist tree structured models which satisfy this set of conditions.

1.6.1 Max-marginals

We start by defining max-marginals since they arise when dealing with marginalization in perturbation models. For two adjacent nodes k and j, we define the max-sum message from j to k,

$$m_{j\to k}(x_k;\gamma) = \max_{x_j} \big\{\tilde{\theta}_{jk}(x_j,x_k;\gamma) + \sum_{i\in N(j)\setminus k} m_{i\to j}(x_j;\gamma)\big\} \quad (1.17)$$

the corresponding maximizing assignment,

$$\hat{x}_{j\to k}(x_k;\gamma) = \arg\max_{x_j} \big\{\tilde{\theta}_{jk}(x_j,x_k;\gamma) + \sum_{i\in N(j)\setminus k} m_{i\to j}(x_j;\gamma)\big\} \quad (1.18)$$

and the resulting max-marginal for node k, $m_k(x_k;\gamma)$, which sums over all the neighbors,

$$m_k(x_k;\gamma) = \sum_{j\in N(k)} m_{j\to k}(x_k;\gamma). \quad (1.19)$$

Furthermore, we use $m_{j\to k}(\gamma)$ to refer to the vector of messages from j to k, whose coordinates are the individual messages $m_{j\to k}(x_k;\gamma)$, and similarly we use $\hat{x}_{j\to k}(\gamma)$ for the vector of maximizing assignments.

Conditioning typically implies comparing differences of messages. To this end, we define for simplicity *normalized* messages and max-marginals by subtracting from each dimension the maximum over the vector of messages:

$$\bar{m}_{j\to k}(x_k,\gamma) = m_{j\to k}(x_k,\gamma) - \max_{x'_k} m_{j\to k}(x'_k;\gamma) \tag{1.20}$$

$$\bar{m}_{k}(x_k,\gamma) = m_{k}(x_k,\gamma) - \max_{x'_k} m_{k}(x'_k;\gamma) \tag{1.21}$$

After normalization, the difference of max-marginals is preserved $\bar{m}_k(x_k;\gamma) - \bar{m}_k(x'_k;\gamma) = m_k(x_k;\gamma) - m_k(x'_k;\gamma), \forall x_k, x'_k \in \mathcal{X}$, and the same is true for individual messages.

Note that the various quantities defined here are random variables induced by the perturbations γ and it makes sense to talk about their pairwise statistical dependency. One possible question is whether the messages $m_{j\to k}(x_k;\gamma)$ or $\bar{m}_{j\to k}(x_k;\gamma)$ are independent of the corresponding maximizing assignments $\hat{x}_{j\to k}(x_k;\gamma)$. Clearly this is true whenever the noise magnitudes are limited such that the maximizing assignments do not depend on the particular noise realizations. Similarly, the independence statement is trivially true whenever the individual messages or the normalized messages are constant with respect to perturbations (i.e. when randomizations "cancel out" regardless of the maximizing assignments). For instance, this is possible when perturbation variables are dependent, like in the proof of Theorem 1.3. One remaining open question is whether there are distributions of perturbations γ for which the statement is more generally true.

In the following we will show how the statistical dependency of max-marginals and maximizing assignments relate to conditional independency in perturbation models.

1.6.2 Conditional Independence

Since low-order perturbations typically give rise to dependencies that go beyond the structure of the potential function, one key question is whether any conditional independencies are maintained.

The first lemma claims that in a three-variable chain $x_1 - x_2 - x_3$, the conditional independence statement $x_1 \perp\!\!\!\perp x_3 | x_2$ holds if for at least one of the two neighbors, the normalized max-marginals are independent of the corresponding maximizing assignments.

Lemma 1.4. *Assume a chain structured model with* 3 *variables* x_1, x_2, x_3,

a randomized potential function of the form $\tilde{\theta}(x) = \tilde{\theta}_{12}(x_1, x_2) + \tilde{\theta}_{23}(x_2, x_3)$ *such that* $\tilde{\theta}_{12} \perp\!\!\!\perp \tilde{\theta}_{23}$*, and the induced perturbation model* $p(x)$*. Then the independence statement* $x_1 \perp\!\!\!\perp x_3 | x_2$ *holds if one of the following statements holds:*

$$\hat{x}_{1\to 2}(x_2, \gamma) \perp\!\!\!\perp \bar{m}_{1\to 2}(x_2', \gamma) \; \forall x_2, x_2' \tag{1.22}$$

$$\hat{x}_{3\to 2}(x_2, \gamma) \perp\!\!\!\perp \bar{m}_{3\to 2}(x_2', \gamma) \; \forall x_2, x_2' \tag{1.23}$$

Proof. When conditioning on $x_2 = \hat{x}_2$ we restrict the perturbations γ to the set $\Gamma(\hat{x}_2)$, defined via: $\mathbf{1}[\gamma \in \Gamma(\hat{x}_2)] = \prod_{x_2 \in \mathcal{X}} \mathbf{1}[m_{1\to 2}(\hat{x}_2; \gamma) + m_{3\to 2}(\hat{x}_2; \gamma) \geq m_{1\to 2}(x_2; \gamma) + m_{3\to 2}(x_2; \gamma)]$. This can be more compactly written via max-marginals: $\mathbf{1}[\gamma \in \Gamma(\hat{x}_2)] = \prod_{x_2 \in \mathcal{X}} \mathbf{1}[m_2(\hat{x}_2; \gamma) - m_2(x_2; \gamma) \geq 0] = \prod_{x_2 \in \mathcal{X}} \mathbf{1}[\bar{m}_2(\hat{x}_2; \gamma) - \bar{m}_2(x_2; \gamma) \geq 0]$.

If condition (1.22) holds, then $\hat{x}_{1\to 2}(\hat{x}_2; \gamma) \perp\!\!\!\perp \bar{m}_2(x_2', \gamma), \forall x_2'$, which implies $\hat{x}_{1\to 2}(\hat{x}_2; \gamma) \perp\!\!\!\perp \bar{m}_2(\hat{x}_2; \gamma) - \bar{m}_2(x_2'; \gamma), \forall x_2'$ and finally $\hat{x}_{1\to 2}(\hat{x}_2; \gamma) \perp\!\!\!\perp \mathbf{1}[\gamma \in \Gamma(\hat{x}_2)]$. Furthermore $\hat{x}_{1\to 2}(\hat{x}_2; \gamma) \perp\!\!\!\perp \hat{x}_{3\to 2}(\hat{x}_2; \gamma)$ from the independence of perturbations across edges and assignments. We can then show that

$$\begin{aligned}
& p(\hat{x}_1, \hat{x}_3 | \hat{x}_2) && (1.24)\\
= \; & \Pr(\hat{x}_1, \hat{x}_3 \in \operatorname*{argmax}_{x_1, x_3} \max_{x_2} \tilde{\theta}(x) | \gamma \in \Gamma(\hat{x}_2)) && (1.25)\\
= \; & \Pr(x_1 \in \hat{x}_{1\to 2}(\hat{x}_2; \gamma) \wedge x_3 \in \hat{x}_{3\to 2}(\hat{x}_2; \gamma) | \gamma \in \Gamma(\hat{x}_2)) && (1.26)\\
= \; & \Pr(x_3 \in \hat{x}_{3\to 2}(\hat{x}_2; \gamma) | \gamma \in \Gamma(\hat{x}_2)) \Pr(x_1 \in \hat{x}_{1\to 2}(\hat{x}_2; \gamma)) && (1.27)\\
= \; & p(\hat{x}_1 | \hat{x}_2) p(\hat{x}_3 | \hat{x}_2) && (1.28)
\end{aligned}$$

□

Intuitively, the independency between messages and the maximizing assignments is used to enforce that at least one of $\hat{x}_1$ or $\hat{x}_3$ is not affected by the joint constraints imposed to ensure that $\hat{x}_2$ is selected through global maximization.

In the following lemma, we show that under the same restrictions, we can condition on a single node by setting the observed variables to their values.

Lemma 1.5. *Assume a chain structured model with* 3 *variables* x_1, x_2, x_3*, a randomized potential function* $\tilde{\theta}(x) = \tilde{\theta}_{12}(x_1, x_2) + \tilde{\theta}_{23}(x_2, x_3)$ *such that* $\tilde{\theta}_{12} \perp\!\!\!\perp \tilde{\theta}_{23}$*, and the induced perturbation model* $p(x)$*.*

If $\hat{x}_{1\to 2}(x_2, \gamma) \perp\!\!\!\perp \bar{m}_{1\to 2}(x_2', \gamma), \forall x_2, x_2'$ *and* $\Gamma(\hat{x}_2) \neq \emptyset$*, then* $\Pr(x_1 = \hat{x}_{1\to 2}(\hat{x}_2; \gamma)) = p(x_1 | \hat{x}_2)$*. In other words, by fixing* x_2 *and perturbing the edge corresponding to* x_1 *only, we obtain the conditional distribution* $p(x_1 | \hat{x}_2)$*.*

Furthermore, if for all $x_1, x_2, x_3, x'_1, x'_2, x'_3$ *we have:*

$$\hat{x}_{1\to 2}(x_2,\gamma) \perp\!\!\!\perp \bar{m}_{1\to 2}(x'_2,\gamma) \text{ and } \Gamma(x_2)\neq\emptyset, \tag{1.29}$$
$$\hat{x}_{2\to 1}(x_1,\gamma) \perp\!\!\!\perp \bar{m}_{2\to 1}(x'_1,\gamma) \text{ and } \Gamma(x_1)\neq\emptyset, \tag{1.30}$$
$$\hat{x}_{3\to 2}(x_2,\gamma) \perp\!\!\!\perp \bar{m}_{3\to 2}(x'_2,\gamma) \text{ and } \Gamma(x_2)\neq\emptyset, \tag{1.31}$$
$$\hat{x}_{2\to 3}(x_3,\gamma) \perp\!\!\!\perp \bar{m}_{2\to 3}(x'_3,\gamma) \text{ and } \Gamma(x_3)\neq\emptyset, \tag{1.32}$$
$$\hat{x}_{2\to\{1,3\}}(x_1,x_3,\gamma) \perp\!\!\!\perp \bar{m}_{2\to\{1,3\}}(x'_1,x'_3,\gamma) \text{ and } \Gamma(x_1,x_3)\neq\emptyset \tag{1.33}$$

then we can condition by plugging in values for any $p(x_\alpha|x_\beta), \alpha\cap\beta=\emptyset$.

Proof. Using the same argument as above, we have that $\hat{x}_{1\to 2}(\hat{x}_2;\gamma) \perp\!\!\!\perp \mathbf{1}[\gamma\in\Gamma(\hat{x}_2)]$, therefore $p(x_1|\hat{x}_2) = \Pr(x_1 = \hat{x}_{1\to 2}(\hat{x}_2;\gamma)|\mathbf{1}[\gamma\in\Gamma(\hat{x}_2)]) = \Pr(x_1 = \hat{x}_{1\to 2}(\hat{x}_2;\gamma))$.

Furthermore, by applying this for possible subsets of variables in the chain $p(x_1|x_2)$, $p(x_2|x_1)$, $p(x_3|x_2)$, $p(x_2|x_3)$, $p(x_2|x_1,x_3)$ we obtain the set of conditions (1.29)-(1.33). □

We can easily extend these results to tree structured models and show that the restrictions on max-product messages provide a feasible method of conditioning on a single variable. One can ask whether there are any trees that satisfy the conditions above at every node and at every subset of nodes. The next lemma provides an example where the conditions hold at every node, but not at pairs of nodes.

Lemma 1.6. *There is a tree structured model for which the conditions of Lemma 1.5 hold for every node symmetrically. In this case, we can condition on every node by plugging in the fixed values.*

Proof. Consider a tree structured graphical model, with binary random variables in $\{-1,1\}$ and with randomized potential function $\tilde{\theta}(x) = \sum_{(i,j)\in E}\tilde{\theta}_{ij}x_ix_j$. If node l is a leaf, then $m_{l\to k}(x_k;\gamma) = \max_{x_l\in\{-1,1\}}\{\tilde{\theta}_{lk}x_lx_k\} = |\tilde{\theta}_{lk}|$. In general, for any node k and any $l\in N(k)$, we have

$$m_{l\to k}(x_k;\gamma) = \sum_{e\in T(l;k)}|\tilde{\theta}_e| \tag{1.34}$$

where $T(l;k)$ denotes the subtree rooted at node l and which does not contain k, $e\in T(l;k)$ is an edge in the subtree. Furthermore, the normalized messages and maximizing assignments are given by

$$\bar{m}_{l\to k}(x_k;\gamma) = 0 \tag{1.35}$$
$$\hat{x}_{l\to k}(x_k;\gamma) = sgn(\tilde{\theta}_{lk}x_k) \tag{1.36}$$

Since the normalized message is always equal to 0, we have $\hat{x}_{l\to k} \perp\!\!\!\perp \bar{m}_{l\to k}, \forall k, \forall l\in N(k)$ and therefore this model satisfies the conditions and

we can do plug-in conditioning for any node k.

However, the two-variable conditions do not hold. Assume $n = 3$ and consider conditioning on x_1, x_3:

$$m_{2\to 1,3}(x_1, x_3; \gamma) = |\tilde{\theta}_{12}x_1 + \tilde{\theta}_{23}x_3| \quad (1.37)$$

$$\bar{m}_{2\to 1,3}(x_1, x_3; \gamma) = |\tilde{\theta}_{12}x_1 + \tilde{\theta}_{23}x_3| - \max_{x_1,x_3} |\tilde{\theta}_{12}x_1 + \tilde{\theta}_{23}x_3| \quad (1.38)$$

$$\hat{x}_{2\to 1,3}(x_1, x_3; \gamma) = sgn(\tilde{\theta}_{12}x_1 + \tilde{\theta}_{23}x_3) \quad (1.39)$$

In this case, $\bar{m}_{2\to 1,3}(\gamma)$ and $\hat{x}_{2\to 1,3}(\gamma)$ will not be independent in general. □

In this section we provided a preliminary analysis of conditioning in perturbation models. In particular, we showed how max-marginals can provide sufficient conditions for conditional independencies with respect to single variables. Unfortunately the methods do not easily extend to conditioning on sets of variables, which remains an open question. Furthermore, we showed examples of perturbations which satisfy the restrictions in Lemma 1.4, which typically involve either the maximizing assignments or the messages to be constant with respect to perturbations. A further question to explore is whether there is a more general characterization of the type of perturbations that satisfy these restrictions.

1.7 Learning Perturbation Models

One of the most distinctive characteristics of perturbation models is that they give rise to dependencies that are not expressed in the base potential function. In the previous chapters we showed that such dependencies arise even when perturbations are independent across the different potential function terms, and across the local assignments within a term. Going a step further, if the perturbations are allowed to be coupled, then we can learn to create and enforce dependencies. This suggests that perturbation models have modeling capacity beyond their base Gibbs' distributions. For example, a tree-structured base model is itself rather restrictive but can be used to induce interactions of all orders in a perturbation setting.

In this section, our goal is to take advantage of this modeling power and learn perturbation models from data. Unlike Gibbs' models, the connection between the structure of the potential function and the properties of the induced distribution is less understood. To this end, we consider complex potential functions equipped with efficient algorithms for computing the maximizing assignment and with expressive dependent perturbations and rely on the learning algorithm to infer the optimal dependency structure.

For the rest of the chapter, we define perturbation models with respect to linear potential functions of the form $\theta(x, w) = w^T \phi(x)$, where w is a vector of parameters and $\phi(x)$ is a vector of features. For instance, for image segmentation, where the prediction is determined by binary variables per pixel location $x = (x_{ij})_{i\in[n],j\in[m]} \in \{0, 1\}^{n\times m}$, a possible feature may check whether neighboring pixels (i, j) and (k, l) are assigned the same class $\phi_{ij,kl}(x_{ij}, x_{kl}) = \mathbf{1}[x_{ij} = x_{kl}]$. In contrast to additive perturbations considered earlier, we define w directly as a random variable. The distribution $p(w; \eta)$ governs the randomization and η are the (hyper-)parameters we aim to learn. This includes the additive case as a special case by simply using $w = w_0 + \gamma$ where w_0 are fixed parameters and γ is a vector of random perturbations.

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

$$\mathcal{P}(\hat{x}; \eta) = \int p(w; \eta)[[\hat{x} = \operatorname*{argmax}_x \theta(x; w)]]dw \tag{1.40}$$

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 continuous hidden variables w. In principle, we could use the EM algorithm resulting in the following iterative updates

$$\eta^{(t+1)} = \operatorname*{argmax}_\eta \sum_{\hat{x}\in\mathcal{S}} E_{w\sim p(w|\hat{x};\eta^{(t)})}\big[\log p(w; \eta)\big] \tag{1.41}$$

Evaluating the expectation requires sampling from the inverse set $\Gamma(\hat{x})$. One way of approaching this issue is to use specialized MCMC algorithms. For instance, (Tarlow et al., 2012) uses a Slice Sampling algorithm which takes advantage of the structure of the problem to avoid fully recomputing the maximizing assignment at every step.

The second approach, which we pursue in this chapter, is to 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

$$\eta^{(t+1)} = \operatorname*{argmax}_\eta \sum_{\hat{x}\in\mathcal{S}} \max_{w\in\Gamma(\hat{x})} \log p(w; \eta) \tag{1.42}$$

While this approach requires a single inner maximization, the problem 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 $x \in dom(\phi)$. However, we will show below that there are many problems of interest for which the inverse set can be described compactly.

1.7.1 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 $\{z : Az \leq b, z \geq 0\}$, 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 (1.42) as a convex program:

$$\max_{w,y} \quad \log p(w;\eta) \tag{1.43}$$

$$s.t. \quad A^T y \geq w, y \geq 0 \tag{1.44}$$

$$y_i = 0, \text{ for } i \in \{i|[A\hat{z}]_i < b\} \tag{1.45}$$

$$[A^T y]_j = w_j, \text{ for } j \in \{j|\hat{z}_j > 0\} \tag{1.46}$$

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 (Ahuja and Orlin, 2001). Also see (Chatalbashev) 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 (1.45-1.46) 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 (Schrijver, 2003). Note that sampling and computing the partition function remain #P-complete (Valiant, 1979) though MCMC-based fully-polynomial approximation schemes exist (Jerrum et al., 2004b). 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 $\sum_{k \in I} z_{kj} = 1, \sum_{k \in J} z_{ik} = 1$ (Schrijver, 2003). 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:

$$\begin{aligned} \max_{w,u,v} \quad & \log p(w; \eta) \\ s.t. \quad & u_i + v_j = w_{ij}, \ (i,j) \in \{(i,j) | \hat{z}_{ij} \neq 0\} \\ & u_i + v_j \leq w_{ij}, \ (i,j) \in \{(i,j) | \hat{z}_{ij} = 0\} \end{aligned}$$

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 (Yang and Ramanan, 2011). 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 $\hat{x}$ (Fromer and Globerson, 2009b). In other words, $\hat{\mu}$ describes $\hat{x}$ in terms of local marginals. Using dual certificates, we can write the inverse problem as:

$$\begin{aligned}
\max_{w} \quad & \log p(w;\eta) \\
s.t. \quad & y_i - \sum_j y'_{i,j;x_i} - \sum_j y''_{j,i;x_i} \geq w_{i;x_i}, \text{for } \hat{\mu}_{i;x_i} = 0 \\
& y_i - \sum_j y'_{i,j;x_i} - \sum_j y''_{j,i;x_i} = w_{i;x_i}, \text{for } \hat{\mu}_{i;x_i} > 0 \\
& y'_{i,j;x_i} + y''_{i,j;x_j} \geq w_{i,j;x_i,x_j}, \text{for } \hat{\mu}_{i,j;x_i,x_j} = 0 \\
& y'_{i,j;x_i} + y''_{i,j;x_j} = w_{i,j;x_i,x_j}, \text{for } \hat{\mu}_{i,j;x_i,x_j} > 0
\end{aligned}$$

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 (Szeliski et al., 2007).

The s-t cut problem can be formulated as the following LP with $m + n$ variables and m constraints:

$$\begin{aligned}
\min_{z} \quad & w^T z \\
s.t. \quad & z_j + z_{S,j} \geq 1 \quad (s,j) \in E \\
& z_j - z_i + z_{i,j} \geq 0 \quad (i,j) \in E \\
& -z_i + z_{i,T} \geq 1 \quad (i,T) \in E
\end{aligned}$$

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 (McDonald et al., 2005). 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 $u \in V \backslash \{r\}$ there is an r-u flow $f^{(u)}$ of value 1 with $f^{(u)}_{ij} \leq z_{ij}$ (Schrijver, 2003). 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.

Example 5. Subset Selection

The subset selection problem appears in machine learning in the context of feature selection, video or text summarization, and others. The prevalence of the problem has lead to various modeling approaches, including budget-based formulations which are typically intractable even for MAP computations, and sub-modular formulations, which are often difficult to sample from. The sub-modular approaches are often optimized in this case

via provable greedy approximations (Gygli et al., 2015) and the perturbation models can be defined as distributions of the (approximate) solution under perturbations of the parameters. In this section we will instead focus on the budget-based approaches and illustrate the inverse LP approach when the formulation is a relaxation.

Consider the task of selecting a fixed number of items (given by a budget B). Specifically, consider the scoring function $\theta(y) = \sum_\alpha \theta_\alpha(y_\alpha)$ where $y_i \in \{0, 1\}$ denotes the absence or presence of an item. In the context of video and text summarization, unary potentials may encode local information, such as the interestingness of the video chunk or sentence, pairwise potentials may encode the similarity between the two items and how far apart they are in the sequence, and so on. See (Gygli et al., 2015) for a range of objectives to consider for video summarization, and (Almeida and Martins, 2013) for text. The goal is to solve $\max_y \theta(y)$ s.t. $\sum_i L_i y_i \leq B$, where L_i is a weight associated with the selected item, e.g. number of frames in the video chunk or number of words in the sentence. Furthermore, we are interested in distributions over subsets, defined as $p(y) \propto \exp(\theta(y))$.

The optimization problem is in general intractable, as it includes the knapsack problem as a special case, and it can be approached by formulating an LP relaxation. For instance, (Almeida and Martins, 2013) use dual-decomposition for optimizing a knapsack objective for text summarization. Their coverage-based summarization model considers M possible sentence topics $(T_m)_{m=1}^M$ with associated relevance scores $w_m \geq 0$ and the goal is to select the subset of sentences that maximizes the overall relevance of the topics covered. Specifically, if $y \in \{0, 1\}^N$ and $u \in \{0, 1\}^M$ are binary vectors denoting the selected sentences and topics respectively, the integer optimization problem to be solved is given by:

$$\max_{u \in \{0,1\}, y \in \{0,1\}} \quad \sum_{m=1}^{M} w_m u_m \tag{1.47}$$

$$s.t. \qquad u_m \leq \sum_{i \in T_m} y_i, \quad \forall m \in \{1 \dots M\} \tag{1.48}$$

$$\sum_{n=1}^{N} L_n y_n \leq B \tag{1.49}$$

After relaxing the integrality constraints and considering the dual, we introduce $(N + 2M + 1)$ new parameters: b and $(s_m)_{m=1}^N$ associated with the budget constraint and topic selection constraints, and $(\alpha_m)_{m=1}^M, (\beta_n)_{n=1}^N$ associated with the constraint of variables being less than 1. Finally, the

optimality conditions lead to the following inverse problem:

$$\min_{w} \quad p(w;\eta) \tag{1.50}$$
$$s.t. \quad s_m + \alpha_m = w_m, \quad \forall m, u_m > 0 \tag{1.51}$$
$$s_m + \alpha_m \geq w_m, \quad \forall m, u_m = 0 \tag{1.52}$$
$$bL_n + \beta_n = \sum_{m:\ n \in T_m} s_m, \quad \forall n, y_n > 0 \tag{1.53}$$
$$bL_n + \beta_n \geq \sum_{m:\ n \in T_m} s_m, \quad \forall n, y_n = 0 \tag{1.54}$$
$$b, s, \alpha, \beta \geq 0 \tag{1.55}$$

1.7.2 Penalty-based Inverse Optimization

The inverse optimization framework provides a clean way of solving the inner maximization in (1.42) 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 (Tsochantaridis et al., 2004) and a similar approach has been explored in (Tarlow et al., 2012). 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.

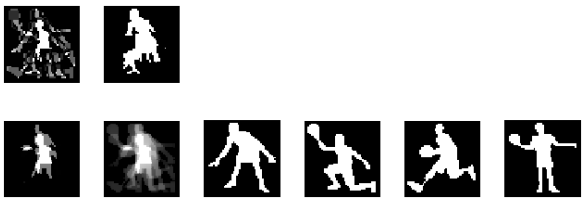

Figure 1.1: First line: max-margin parameters and resulting segmentation, second line: the mean of the perturbation parameters, the average segmentation and the four images with the highest count.

1.8 Empirical Results

We conclude the chapter by presenting experiments demonstrating that perturbation models capture dependencies above and beyond the structure of the potential function. The first experiment explores an image segmentation task and illustrates the duality approach for learning perturbation models. While the potential function is formed of local pairwise potentials and implies long range conditional independencies, the experiment suggests that in the learned perturbation model various long range independencies do not hold. The second experiment shows an application of learning perturbation models in the context of image matching.

1.8.1 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 each example, initialized with the point estimate from the previous hard-EM iteration. Since the setting is

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

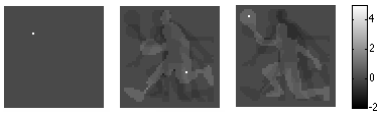

Figure 1.2: Correlations between a reference pixel (white) and the rest, as captured by the covariance matrix of the perturbation distribution. We show a pixel that is always off (so no correlations) and two pixels that are activated on different poses.

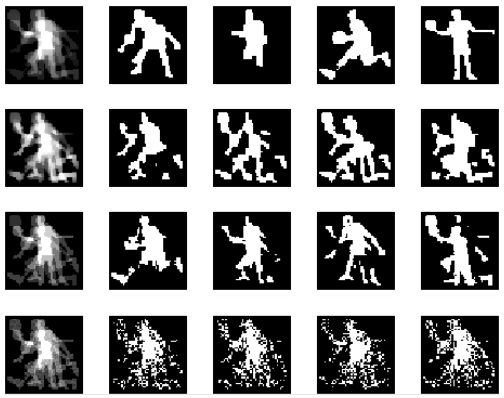

Figure 1.3: The average segmentation and samples from four models, one per line: perturbation model where the perturbations have unrestricted vs. diagonal covariance matrix and multivariate gaussian model with unrestricted vs. diagonal covariance matrix.

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 1.1, 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 1.2. 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 1.3, 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 1.3, 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.

1.8.2 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 (Yang and Ramanan, 2011) 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 (Volkovs and Zemel, 2012), $\theta(I, I', \pi; w) = \sum_{i,j} w^T(\psi(I, i) - \psi(I', j))^2$. 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 1.4 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.

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

Figure 1.4: 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.

1.9 Perturbation Models and Stability

In the previous sections we showed how to estimate perturbation models from data and demonstrated their extended modeling power. To this end, we focused on base models where the MAP assignment can be evaluated efficiently even if the marginals (or the partition function) of the base Gibbs model is not feasible. Such models remain learnable within the perturbation framework, enriched by induced longer range dependencies.

The situation changes when the family of base potential functions no longer permits efficient MAP assignments. For instance, in section 1.7.1 we indicate how approximations can be used with inverse optimization. More generally, tractability may arise as a by-product of learning perturbation models. Indeed, while randomization is needed to introduce diversity in samples, maximizing the likelihood of the correct assignment also serves to carve out stable assignments. Stability, on the other hand, can be related to tractability. We start by describing various notions of stability and their relationship to the hardness of inference calculations.

The complex models we consider here are common in applications in natural language processing, computer vision and bioinformatics that involve clusters, parse trees, or arrangements. As a result, much of the work in structured prediction has focused on designing heuristics for inference, such as loopy belief propagation (Murphy et al., 1999), tree reweighed message passing (Wainwright et al., 2005), local search algorithms (Zhang et al., 2014) or convex relaxations (Koo et al., 2010b), and empirical results show that these methods are often successful in recovering the correct (target) solution (Koo et al., 2010b; Rush et al., 2010; Zhang et al., 2014). This suggests that the instances encountered during inference are much easier than indicated by their complexity class.

The success of the heuristics can be attributed to the additional structural properties that are present in the typical instances. For instance, if the target

solution stands out amongst all other solutions in some manner, than we expect heuristic approaches to discover it in polynomial time.

In theoretical computer science, the relevant work has focused on identifying the interesting structural properties which can be exploited to design specialized new algorithms or to prove the correctness of current heuristics. Such properties include Bilu-Linial stability (Bilu et al., 2012; Awasthi et al., 2012; Bilu and Linial, 2012; Makarychev et al., 2014), approximation stability (Balcan and Liang, 2012), weak-deletion stability (Awasthi et al., 2012, 2010), and so on. For instance, the notion of Bilu-Linial γ-stability specifies that the optimal solution does not change upon multiplicative perturbations of the parameters of magnitude at most γ and in this case (Makarychev et al., 2014) showed that Max-Cut is tractable whenever $\gamma \geq \sqrt{n} \log \log n$ for some constant c.

In structured prediction, the additional properties that trigger the success of approximate inference procedures can be attributed to the learning algorithms used to estimate the parameters. For example, one of the common learning strategies is to maximize the margin between the target solution and potential candidates: $\theta(\hat{x}) - \theta(x) \geq \gamma\Delta(\hat{x}, x), \forall x$, where Δ is a distance measure between assignments, allowing a closer margin between similar assignments. This notion of stability (margin stability) has been empirically proven to produce tractable instances under various approximate inference algorithms (Finley and Joachims, 2008). Also, from the theoretical perspective, one can relate the notion of margin (additive) stability to the multiplicative stability mentioned above to provide weak guarantees, which suggests that explicitly enforcing the saliency of target solutions brings computational benefits for inference.

Even more concretely, the additive margins can be related to the empirical success of various linear programming relaxations approaches in machine learning. For instance, considering scoring functions of the form $\sum_\alpha \theta_\alpha(x_\alpha)$ on binary assignments $x \in \{0, 1\}^n$, the dual decomposition algorithm (Koo et al., 2010b; Sontag et al., 2011) has been successfully used for parsing with high order interactions, despite the theoretical intractability of the problem. To illustrate this, consider the optimality conditions for the resulting linear program. When most of the local potentials agree on the maximizing assignment, the relaxation is tight:

Lemma 1.7. *Assuming that a $(1 - \delta)$ fraction of the components support the correct solution with a margin γ (i.e. for most α, $\theta_\alpha(x^*_\alpha) > \theta_\alpha(x_\alpha) + \gamma$), and the remaining δ fraction do not object by more than M (i.e. $\theta_\alpha(x^*_\alpha) > \theta_\alpha(x_\alpha) - M$) and $\delta \leq \frac{\gamma}{\gamma+M}$, then the dual-decomposition algorithm returns the correct solution.*

Proof. Consider the binary structured prediction problem where the maximizing assignment is given by $\hat{x} = \arg\max_x \theta(x)$. We start by rewriting it as $\hat{x} = \arg\max_{x=x'} \sum_{i=1}^n \theta_i(x_i) + \sum_\alpha \theta_\alpha(x'_\alpha)$, where we added constant unary potentials $\theta_i(x_i)$ for $i \in \{1 \dots n\}$. Solving the optimization problem via dual decomposition involves computing

$$\begin{aligned}
\hat{\delta} &= \arg\min_\delta \big\{ \max_x (\sum_i (\theta_i(x_i) - \sum_\alpha \delta_{i,\alpha}(x_i))) + \sum_\alpha \max_{x'_\alpha} (\theta_\alpha(x'_\alpha) + \sum_i \delta_{i,\alpha}(x'_i)) \big\} \\
\hat{x} &= \arg\max_x \big\{ \sum_i (\theta_i(x_i) - \sum_\alpha \hat{\delta}_{i,\alpha}(x_i)) \big\} \\
\hat{x}'_\alpha &= \arg\max_{x'_\alpha} \big\{ \theta_\alpha(x'_\alpha) + \sum_i \hat{\delta}_{i,\alpha}(x'_i) \big\}
\end{aligned}$$

To show that a target assignment x^* is optimal, we find a dual witness δ^* such that: $\max_x (\sum_i (\theta_i(x_i) - \sum_\alpha \delta^*_{i,\alpha}(x_i))) + \sum_\alpha \max_{x'_\alpha} (\theta_\alpha(x'_\alpha) + \sum_i \delta^*_{i,\alpha}(x'_i)) \leq \sum_i \theta_i(x^*_i) + \sum_\alpha \theta_\alpha(x^*_\alpha)$.

Define $\delta^*_{i,\alpha}(x^*_i) = 0$ and $\delta^*_{i,\alpha}(1 - x^*_i) = \min_{x''_\alpha, \Delta(x^*_\alpha, x''_\alpha) > 0} \frac{\theta_\alpha(x^*_\alpha) - \theta_\alpha(x''_\alpha)}{\Delta(x^*_\alpha, x''_\alpha)}$, where $\Delta(\cdot, \cdot)$ counts the number of dimension where the assignments disagree. Specifically, we design the dual witness δ^* such that it enforces local optimality of the target solution by increasing/decreasing the weight of the alternative local solutions.

With this choice of dual variables and an arbitrary assignment x'_α, we have: $\theta_\alpha(x'_\alpha) + \sum_i \delta^*_{i,\alpha}(x'_i) = \theta_\alpha(x'_\alpha) + \sum_{i, x'_i \neq x^*_i} \min_{x''_\alpha, \Delta(x^*_\alpha, x''_\alpha) > 0} \frac{\theta_\alpha(x^*_\alpha) - \theta_\alpha(x''_\alpha)}{\Delta(x^*_\alpha, x''_\alpha)}$ $\leq \theta_\alpha(x'_\alpha) + \Delta(x^*_\alpha, x'_\alpha) \frac{\theta_\alpha(x^*_\alpha) - \theta_\alpha(x'_\alpha)}{\Delta(x^*_\alpha, x'_\alpha)} = \theta_\alpha(x^*_\alpha)$. Therefore, all the modified local potentials select the target assignment via maximization and $\max_{x'_\alpha} \{\theta_\alpha(x'_\alpha) + \sum_i \delta^*_{i,\alpha}(x'_i)\} = \theta_\alpha(x^*_\alpha)$.

To conclude the proof we need to show that $\max_x \{\sum_i (\theta_i(x_i) - \sum_\alpha \delta^*_{i,\alpha}(x_i))\} \leq \sum_i \theta_i(x^*)$. We have:

$$\begin{aligned}
\sum_i (\theta_i(x_i) - \sum_\alpha \delta^*_{i,\alpha}(x_i)) &= \sum_i \theta_i(x_i) - \sum_{i, x_i \neq x^*_i} \sum_\alpha \min_{x''_\alpha, \Delta(x^*_\alpha, x''_\alpha) > 0} \frac{\theta_\alpha(x^*_\alpha) - \theta_\alpha(x''_\alpha)}{\Delta(x^*_\alpha, x''_\alpha)} \\
&\leq \sum_i \theta_i(x_i) - ((1 - \delta)\gamma - \delta M)
\end{aligned}$$

where we used that $\theta_\alpha(x^*_\alpha) - \theta_\alpha(x''_\alpha) \geq \gamma$ for a $(1 - \delta)$ fraction of the local potentials and $\theta_\alpha(x^*_\alpha) - \theta_\alpha(x''_\alpha) \geq -M$ for the rest.

If $((1 - \delta)\gamma - \delta M \geq \max_x \sum_i (\theta_i(x_i) - \theta_i(x^*))$, then the target solution is optimal for the dual decomposition algorithm. Since θ_i were introduced as constant local potentials, we have that $\delta \leq \frac{\gamma}{\gamma + M}$ is sufficient to imply the optimality of the target solution.

□

The observations in this section argue for enforcing stability with respect to perturbations of the parameters. In fact, dual-decomposition-based inference has been successfully applied in conjunction with simple learning algorithms which encourage local assignments to be consistent with the overall solution (Koo et al., 2010b).

Learning perturbation models is inherently tied to stability. Maximizing the probability that a perturbation model realizes a given answer also encourages the answer to be stable, carrying tractability benefits. Indeed, perturbation models can be tailored to achieve various notions of stability by designing appropriate (e.g. multiplicative) perturbations. Such variations can remain tractable even if the base model (as a class) is not.

1.10 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 (Koller and Friedman, 2009; Wainwright and Jordan, 2008). The Gibbs distribution as well as its Markov properties can be realized from the statistics of high dimensional random MAP perturbations with the Gumbel distribution (see Theorem 1.1), (Papandreou and Yuille, 2011; Tarlow et al., 2012; Hazan and Jaakkola, 2012; Hazan et al., 2013). For comprehensive introduction to extreme value statistics we refer the reader to (Kotz and Nadarajah, 2000).

Recent work (Papandreou and Yuille, 2010, 2011; Tarlow et al., 2012) explores the different aspects of low dimensional MAP perturbation models. Papandreou et al. (Papandreou and Yuille, 2010) describe sampling from the Gaussian distribution with random Gaussian perturbations. Later (Papandreou and Yuille, 2011), 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. (Tarlow et al., 2012) 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 cut-

ting 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 (Yanover and Weiss, 2004; Fromer and Globerson, 2009a; Batra, 2012; Mezuman et al., 2013; Batra et al., 2012; Guzman-Rivera et al., 2012). Finding the 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 (Kulesza and Taskar, 2012). This learning approach concerns problems that can be described by determinants while our approach is based on MRF potentials.

1.11 References

R. K. Ahuja and J. B. Orlin. Inverse optimization. In *Operations Research*, 2001.

M. B. Almeida and A. F. Martins. Fast and robust compressive summarization with dual decomposition and multi-task learning. In *ACL (1)*, pages 196–206, 2013.

P. Awasthi, A. Blum, and O. Sheffet. Clustering under natural stability assumptions. 2010.

P. Awasthi, A. Blum, and O. Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1):49–54, 2012.

M. F. Balcan and Y. Liang. Clustering under perturbation resilience. In *Automata, Languages, and Programming*, pages 63–74. Springer, 2012.

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

D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In *ECCV*, 2012.

Y. Bilu and N. Linial. Are stable instances easy? *Combinatorics, Probability and Computing*, 21(05):643–660, 2012.

Y. Bilu, A. Daniely, N. Linial, and M. Saks. On the practically interesting instances of maxcut. *arXiv preprint arXiv:1205.4893*, 2012.

V. Chatalbashev. Inverse convex optimization.

H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015.

T. Finley and T. Joachims. Training structural svms when exact inference is intractable. In *Proceedings of the 25th international conference on Machine learning*, pages 304–311. ACM, 2008.

M. Fromer and A. Globerson. An lp view of the m-best map problem. *Advances in Neural Information Processing Systems (NIPS)*, 22:567–575, 2009a.

M. Fromer and A. Globerson. An lp view of the m-best map problem. *Advances in Neural Information Processing Systems (NIPS)*, 2009b.

A. Gane, T. Hazan, and T. Jaakkola. Learning with maximum a-posteriori perturbation models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 1984.

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

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

A. Guzman-Rivera, P. Kohli, and D. Batra. Faster training of structural svms with diverse m-best cutting-planes. In *Discrete Optimization in Machine Learning Workshop (DISCML-NIPS)*, 2012.

M. Gygli, H. Grabner, and L. Van Gool. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3090–3098, 2015.

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

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

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*, 2013.

M. Huber. A bounding chain for swendsen-wang. *Random Structures and Algorithms*, 2003.

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

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, 2004a.

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, 2004b.

J. H. Kappes, P. Swoboda, B. Savchynskyy, T. Hazan, and C. Schnörr. Probabilistic correlation clustering and image partitioning using perturbed multicuts. In *Scale Space and Variational Methods in Computer Vision*, pages 231–242. Springer, 2015.

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

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

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

T. Koo, A. M. Rush, M. Collins, T. Jaakkola, and D. Sontag. Dual decomposition

for parsing with non-projective head automata. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1288–1298. Association for Computational Linguistics, 2010b.

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

A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 2012.

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.

K. Makarychev, Y. Makarychev, and A. Vijayaraghavan. Bilu-linial stable instances of max cut and minimum multiway cut. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 890–906. SIAM, 2014.

R. McDonald and G. Satta. On the complexity of non-projective data-driven dependency parsing. In *Proceedings of the 10th International Conference on Parsing Technologies*, pages 121–132. Association for Computational Linguistics, 2007.

R. McDonald, F. Pereira, K. Ribarov, and J. Hajič. Non-projective dependency parsing using spanning tree algorithms. In *EMNLP*, 2005.

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.

K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999.

S. Nowozin and C. H. Lampert. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4): 185–365, 2011.

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

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

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

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

L. Rabiner and B.-H. Juang. Fundamentals of speech recognition. 1993.

A. M. Rush, D. Sontag, M. Collins, and T. Jaakkola. On dual decomposition and linear programming relaxations for natural language processing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1–11. Association for Computational Linguistics, 2010.

A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*. Springer, 2003.

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.

D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. *Optimization for Machine Learning*, 1:219–254, 2011.

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.

D. Tarlow, R. Adams, and R. Zemel. Randomized optimum models for structured prediction. In *AISTATS*, 2012.

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.

L. Valiant. The complexity of computing the permanent. *Theoretical computer science*, 1979.

M. Volkovs and R. S. Zemel. Efficient sampling for bipartite matching problems. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 2008.

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.

J. Wang and R. Swendsen. Nonuniversal critical dynamics in monte carlo simulations. *Physical review letters*, 1987.

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

Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

C. Yanover and Y. Weiss. Finding the m most probable configurations using loopy belief propagation. *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Y. Zhang, T. Lei, R. Barzilay, and T. Jaakkola. Greed is good if randomized: New inference for dependency parsing. *EMNLP*, 2014.