Journal of Machine Learning Research 20 (2019) 1-34 Submitted 9/17; Revised 10/18; Published 2/19

Train and Test Tightness of LP Relaxations in Structured Prediction

Ofer Meshi MESHI@GOOGLE.COM
Google

Ben London BLONDON@AMAZON.COM
Amazon

Adrian Weller ADRIAN.WELLER@ENG.CAM.AC.UK
University of Cambridge and The Alan Turing Institute

David Sontag DSONTAG@CSAIL.MIT.EDU
MIT CSAIL

Editor: Ivan Titov

Abstract

Structured prediction is used in areas including computer vision and natural language processing to predict structured outputs such as segmentations or parse trees. In these settings, prediction is performed by MAP inference or, equivalently, by solving an integer linear program. Because of the complex scoring functions required to obtain accurate predictions, both learning and inference typically require the use of approximate solvers. We propose a theoretical explanation for the striking observation that approximations based on linear programming (LP) relaxations are often tight (exact) on real-world instances. In particular, we show that learning with LP relaxed inference encourages integrality of training instances, and that this training tightness generalizes to test data.

1. Introduction

Many applications of machine learning can be formulated as prediction problems over structured output spaces (Bakir et al., 2007; Nowozin et al., 2014). In such problems output variables are predicted *jointly* in order to take into account mutual dependencies between them, such as high-order correlations or structural constraints (e.g., matchings or spanning trees). Unfortunately, the improved expressive power of these models comes at a computational cost, and indeed, exact prediction and learning become NP-hard in general. Despite this worst-case intractability, efficient approximations often achieve very good performance in practice. In particular, one type of approximation which has proved effective in many applications is based on *linear programming (LP) relaxation*. In this approach the prediction problem is first cast as an integer LP (ILP), and then the integrality constraints are relaxed to obtain a tractable program. In addition to achieving high prediction accuracy, it has been observed that LP relaxations are often *tight* in practice. That is, the solution to the relaxed program happens to be optimal for the original hard problem (i.e., an *integral* solution is found). This is particularly surprising since the LPs have complex scoring functions that are not constrained to be from any tractable family. It

has been an interesting open question to understand why these real-world instances behave so differently from the theoretical worst case.

This paper addresses this question and aims to provide a theoretical explanation for the frequent tightness of LP relaxations in the context of structured prediction. In particular, we show that an *approximate* training objective, although designed to produce accurate predictors, also induces tightness of the LP relaxation as a byproduct. Interestingly, our analysis also suggests that exact training may have the opposite effect. To explain tightness on future (i.e., test) instances, we prove several generalization bounds relating average tightness on the training data to expected tightness with respect to the generating distribution. Our bounds imply that if many predictions on training instances are tight, then predictions on test instances are also likely to be tight. Moreover, if predictions on training instances are *close to* integral solutions (in terms of L1 distance), then predictions on test instances will likely be similarly close to integral solutions. Our results help to understand previous empirical findings, and to our knowledge provide the first theoretical justification for the widespread success of LP relaxations for structured prediction in settings where the training data is not (algorithmically) separable.

2. Background

In this section we review the formulation of the structured prediction problem, its LP relaxation, and the associated learning problem. Consider a prediction task where the goal is to map a real-valued input vector x to a discrete output vector $y = (y_1, \ldots, y_n)$. In this work we focus on a simple class of models based on linear classifiers.[1] Particularly, in this setting prediction is performed via a linear discriminant rule: $y(x; w) = \arg\max_{y'} w^\top \phi(x, y')$, where $\phi(x, y) \in \mathbb{R}^d$ is a function mapping input-output pairs to feature vectors, and $w \in \mathbb{R}^d$ is the corresponding weight vector. Since the output space is often huge (exponential in n), it will generally be intractable to maximize over all possible outputs.

In many applications the score function has a particular structure. Specifically, we will assume that the score decomposes as a sum of simpler score functions: $w^\top \phi(x, y) = \sum_c w_c^\top \phi_c(x, y_c)$, where y_c is an assignment to a (non-exclusive) subset of the variables. For example, it is common to use such a decomposition that assigns scores to single and pairs of output variables corresponding to nodes and edges of a graph G: $w^\top \phi(x, y) = \sum_{i \in V(G)} w_i^\top \phi_i(x, y_i) + \sum_{ij \in E(G)} w_{ij}^\top \phi_{ij}(x, y_i, y_j)$. Viewing this as a function of y, we can write the prediction problem as:

$$\max_y \sum_c \theta_c(y_c; x, w) \ , \tag{1}$$

where $\theta_c(y_c; x, w) = w_c^\top \phi_c(x, y_c)$ (we will sometimes omit the dependence on x and w in the sequel).

Due to its combinatorial nature, the prediction problem is generally NP-hard, but fortunately, efficient approximations have been proposed. Here we will be particularly interested in approximations based on LP relaxations. We begin by formulating prediction

1. Most of our results also hold more generally for non-linear models (e.g., deep neural factors).

as the following ILP:[2]

$$\max_{\substack{\mu\in\mathcal{M}_L\\ \mu\in\{0,1\}^q}} \sum_c \sum_{y_c} \mu_c(y_c)\theta_c(y_c) + \sum_i \sum_{y_i} \mu_i(y_i)\theta_i(y_i) \quad = \theta^\top \mu \ , \tag{2}$$

$$\text{where } \mathcal{M}_L = \left\{ \mu \geq 0 : \begin{array}{ll} \sum_{y_{c\setminus i}} \mu_c(y_c) = \mu_i(y_i) & \forall c, i \in c, y_i \\ \sum_{y_i} \mu_i(y_i) = 1 & \forall i \end{array} \right\}.$$

Here, $\mu_c(y_c)$ is an indicator variable for a factor c and local assignment y_c, and q is the total number of factor assignments (dimension of μ). The set $\mathcal{M}_L$ is known as the local marginal polytope (Wainwright and Jordan, 2008), and $\sum_{y_{c\setminus i}} \mu_c(y_c)$ is the marginalization of μ_c w.r.t. variable i. First, notice that there is a one-to-one correspondence between feasible μ's and assignments y's, which is obtained by setting μ to indicators over local assignments (y_c and y_i) consistent with y. Second, while solving ILPs is NP-hard in general, it is easy to obtain a tractable program by relaxing the integrality constraints ($\mu \in \{0,1\}^q$), which may introduce fractional solutions to the LP. This relaxation, sometimes called the *basic linear programming relaxation* (Thapper and Živný, 2012), is the first level of the Sherali-Adams hierarchy (Sherali and Adams, 1990). This hierarchy provides successively tighter LP relaxations of an ILP. Notice that since the relaxed program is obtained by removing constraints from the original problem, its optimal value upper bounds the ILP optimum. Finally, we note that most of our results below also hold for cases where more complex constraints than those in Eq. (2) are used. For example, some assignments may be forbidden since they do not correspond to feasible global structures such as spanning trees (e.g., Martins et al., 2009b; Koo et al., 2010).

In order to achieve high prediction accuracy, the parameters w are learned from training data. In this supervised learning setting, the model is fit to labeled examples $\{(x^{(m)}, y^{(m)})\}_{m=1}^M$, where the goodness of fit is measured by a task-specific loss $\Delta(y(x^{(m)}; w), y^{(m)})$. We assume that $\Delta(y, y') \geq 0$ for all y, y', and that $\Delta(y, y) = 0$. For example, a commonly used task-loss is the Hamming distance: $\Delta_{\text{Hamming}}(y, y') = \frac{1}{n} \sum_i \mathbb{1}[y_i \neq y'_i]$.

In the *structured SVM* (SSVM) framework (Taskar et al., 2003; Tsochantaridis et al., 2004), the empirical risk is upper bounded by a convex surrogate called the structured hinge loss, which yields the training objective:

$$\min_w \ \sum_m \max_y \left[w^\top \Big(\phi(x^{(m)}, y) - \phi(x^{(m)}, y^{(m)}) \Big) + \Delta(y, y^{(m)}) \right] \ . \tag{3}$$

For brevity, we have omitted the standard regularization term from Eq. (3), however, all of our results below in sections 4—6 still hold with regularization.[3] The objective in Eq. (3) is a convex function of w and hence can be optimized in various ways. But, notice that the objective includes a maximization over outputs y for each training example. This loss-augmented prediction task needs to be solved repeatedly during training (e.g., to evaluate subgradients), which makes training intractable in general (see also Sontag et al., 2010). Similar to prediction, LP relaxation can be applied to the structured loss (Taskar et al.,

2. For convenience we introduce singleton factors θ_i, which could be set to 0 if needed.
3. In particular, our bounds apply to the maximization over y in Eq. (3), so still hold when regularization w.r.t. w is added.

2003; Kulesza and Pereira, 2007), which yields the relaxed training objective:

$$\min_{w} \quad \sum_{m} \max_{\mu \in \mathcal{M}_L} \left[\theta_m^\top (\mu - \mu_m) + {\ell_m}^\top \mu \right] \ , \tag{4}$$

where $\theta_m \in \mathbb{R}^q$ is a score vector in which each entry represents $w_c^\top \phi_c(x^{(m)}, y_c)$ for some c and y_c, and μ_m is the integral vector corresponding to $y^{(m)}$. Assuming that the task-loss decomposes as the model score, $\Delta(y, y') = \sum_c \Delta_c(y_c, y'_c)$, we define the vector $\ell_m \in \mathbb{R}^q$ with entries $\ell_{m,c,y_c} = \Delta_c(y_c^{(m)}, y_c)$ for each value y_c. Notice that ℓ has the same dimension as μ, and we can define $\Delta(y^{(m)}, y) = \sum_c \sum_{y'_c} \Delta_c(y_c^{(m)}, y'_c)\mu_c(y'_c) = {\ell_m}^\top \mu$, where μ is the vector of indicators corresponding to y (i.e., $\mu_c(y'_c) = \mathbb{1}\{y'_c = y_c\}$). With this definition, ℓ generalizes Δ to any $\mu \in \mathcal{M}_L$.

After reviewing related work in Section 3, we propose a theoretical justification for the observed tightness of LP relaxations for structured prediction. To this end, we make two complementary arguments: in Section 4 we argue that optimizing the relaxed training objective of Eq. (4) also has the effect of encouraging tightness of training instances; then, in sections 5 and 6 we show that tightness generalizes from train to test data.

3. Related Work

Many structured prediction problems can be expressed as ILPs (Roth and Yih, 2005; Martins et al., 2009a; Rush et al., 2010). Despite being NP-hard in general (Roth, 1996; Shimony, 1994), various effective approximations have been proposed. These approximations include search-based methods (Daumé III et al., 2009; Zhang et al., 2014) and natural LP relaxations to the hard ILPs (Schlesinger, 1976; Koster et al., 1998; Chekuri et al., 2004; Wainwright et al., 2005). Tightness of LP relaxations for special classes of problems has been studied extensively in recent years and has been demonstrated by restricting either the structure of the model or its score function. For example, the pairwise LP relaxation is known to be tight for tree-structured models or for supermodular scores (see, e.g., Wainwright and Jordan, 2008; Thapper and Živný, 2012); certain stability conditions guarantee tightness of LP relaxations for Ferromagnetic Potts models (Lang et al., 2018); and the cycle relaxation (for binary pairwise models) is known to be tight both for planar Ising models with no external field (Barahona, 1993) and for almost balanced models (Weller et al., 2016). Hybrid conditions, combining structure and score, by forbidding signed minors have recently been shown to also guarantee tight relaxations (Rowland et al., 2017; Weller, 2016). To facilitate efficient prediction, one could restrict the model class to be tractable. For example, Taskar et al. (2004) learn supermodular scores, and Meshi et al. (2013) learn tree structures.

However, the sufficient conditions mentioned above are by no means necessary, and indeed, many score functions that are useful in practice do not satisfy them but still produce integral solutions (Roth and Yih, 2004; Lacoste-Julien et al., 2006; Sontag et al., 2008; Finley and Joachims, 2008; Martins et al., 2009b; Koo et al., 2010). For example, Martins et al. (2009b) showed that predictors that are learned with LP relaxations yield tight LPs for 92.88% of the test data on a dependency parsing problem (see Table 2 therein). Koo et al. (2010) observed similar behavior for dependency parsing on a number of languages, as can

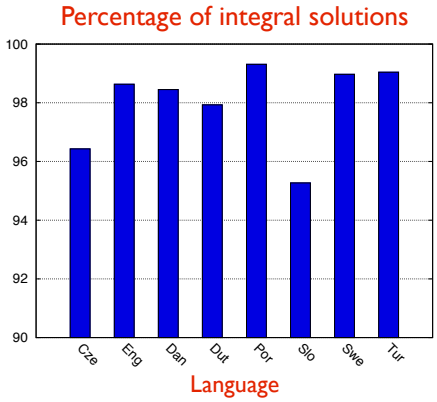

Figure 1: Percentage of integral solutions for dependency parsing from Koo et al. (2010).

be seen in Fig. 1.[4] Lacoste-Julien et al. (2006) found that about 80% of test instances had tight LP relaxations in a quadratic assignment formulation for word alignment, with only 0.2% fractional values overall. The same phenomenon has been observed for a multi-label classification task, where test integrality reached 100% (Finley and Joachims, 2008, Table 3).

Learning structured output predictors from labeled data was proposed in various forms by Collins (2002); Taskar et al. (2003); Tsochantaridis et al. (2004). These formulations generalize training methods for binary classifiers, such as the Perceptron algorithm and support vector machines (SVMs), to the case of structured outputs. The learning algorithms repeatedly perform prediction, necessitating the use of approximate inference within training as well as at test time. A common approach, introduced right at the inception of structured SVMs by Taskar et al. (2003), is to use LP relaxations for this purpose.

The closest work to ours is by Kulesza and Pereira (2007), which showed that not all approximations are equally good, and that it is important to match the inference algorithms used at train and test time. The authors defined the concept of *algorithmic separability* which refers to the case where an approximate inference algorithm achieves zero loss on a data set. The authors studied the use of LP relaxations for structured learning, giving generalization bounds for the true risk of LP-based prediction. However, since the generalization bounds in Kulesza and Pereira (2007) are focused on prediction *accuracy*, the only settings in which tightness on test instances can be guaranteed are when the training data is algorithmically separable, which is seldom the case in real-world structured prediction tasks (the models are far from perfect). In contrast, our paper's main result (Theorem 1), guarantees that the expected fraction of test instances for which an LP relaxation is tight is close to that which was achieved on training data. This then allows us to talk about the generalization of *computation*. For example, suppose one uses LP relaxation-based algorithms that iteratively tighten the relaxation, such as Sontag and Jaakkola (2008); Sontag et al. (2008), and observes that 20% of the instances in the training data are integral using the basic relaxation and that after tightening the remaining 80% are now integral too. Our generalization bound then guarantees that approximately the same ratio will hold at test time (assuming sufficient training data).

4. Kindly provided by the authors.

Finley and Joachims (2008) also studied the effect of various approximate inference methods in the context of structured prediction. Their theoretical and empirical results support the superiority of LP relaxations in this setting. Martins et al. (2009b) established conditions which guarantee algorithmic separability for LP relaxed training, and derived risk bounds for a learning algorithm which uses a combination of exact and relaxed inference.

Finally, Globerson et al. (2015) recently studied the performance of structured predictors for 2D grid graphs with binary labels from an information-theoretic perspective. They proved lower bounds on the minimum achievable expected Hamming error in this setting, and proposed a polynomial-time algorithm that achieves this error. Our work is different since we focus on LP relaxations as an approximation algorithm, we handle a general form of the problem without making any assumptions on the model or error measure (except score decomposition), and we concentrate solely on the computational aspects while ignoring any accuracy concerns.

4. Tightness at Training

In this section we show that the *relaxed* training objective in Eq. (4), although designed to achieve high accuracy, also induces tightness of the underlying LP relaxation. In fact, although we focus here on the basic LP relaxation (first-level), the results below hold for higher-level LP relaxations as well.[5]

In order to simplify notation we focus on a single training instance and drop the index m. Denote the solutions to the relaxed and integer LPs as:

$$\mu_L \in \arg\max_{\mu \in \mathcal{M}_L} \theta^\top \mu \qquad\qquad \mu_I \in \arg\max_{\substack{\mu \in \mathcal{M}_L \\ \mu \in \{0,1\}^q}} \theta^\top \mu \ , \tag{5}$$

respectively. Also, let μ_T be the integral vector corresponding to the ground-truth output $y^{(m)}$. Now consider the following decomposition:

$$\underset{\text{relaxed-hinge}}{\theta^\top(\mu_L - \mu_T)} = \underset{\text{integrality gap}}{\theta^\top(\mu_L - \mu_I)} + \underset{\text{exact-hinge}}{\theta^\top(\mu_I - \mu_T)} \tag{6}$$

This equality states that the difference in scores between the relaxed optimum and ground-truth (*relaxed-hinge*) can be written as a sum of the *integrality gap* and the difference in scores between the exact optimum and the ground-truth (*exact-hinge*); notice that all three terms are non-negative. This simple decomposition has several interesting implications.

First, we can immediately derive the following bound on the integrality gap:

$$\begin{aligned}
\theta^\top(\mu_L - \mu_I) &= \theta^\top(\mu_L - \mu_T) - \theta^\top(\mu_I - \mu_T) && (7)\\
&\leq \theta^\top(\mu_L - \mu_T) && (8)\\
&\leq \theta^\top(\mu_L - \mu_T) + \ell^\top \mu_L && (9)\\
&\leq \max_{\mu \in \mathcal{M}_L} \left(\theta^\top(\mu - \mu_T) + \ell^\top \mu\right), && (10)
\end{aligned}$$

where Eq. (10) is precisely the relaxed training objective from Eq. (4). Therefore, optimizing the approximate training objective of Eq. (4) minimizes an upper bound on the integrality

5. Similar analysis for SDP and quadratic relaxations is left as future work—see recent work by Lê-Huu and Paragios (2018).

gap. Hence, driving down the approximate objective may also reduce the integrality gap of training instances, although in Section 4.1 we also study cases where loose bounds can lead to non-zero integrality gap. One case where the integrality gap becomes zero is when the data is algorithmically separable (i.e., $\mu_L = \mu_T$, so Eq. (8) equals 0). In this case the relaxed-hinge term vanishes (the exact-hinge must also vanish), and training integrality is assured.

Second, Eq. (6) holds for *any integral* μ, and not just the ground-truth μ_T. In other words, the only property of μ_T used here is its integrality. Indeed, in Section 7 we verify empirically that training a model using *random labels* still attains the same level of tightness as training with the ground-truth labels.[6] On the other hand, accuracy drops drastically, as expected. This analysis suggests that *tightness is not coupled with accuracy* of the predictor. Finley and Joachims (2008) explained tightness of LP relaxations by noting that fractional solutions always incur a loss during training. Our analysis suggests an alternative explanation, emphasizing the difference in scores (Eq. (7)) rather than the loss, and decoupling tightness from accuracy.

Third, the results above still hold in the presence of global constraints. Often such constraints can be expressed as linear inequalities, so in this case the local polytope $\mathcal{M}_L$ can be redefined by adding these constraints to those in Eq. (2) to form a tighter polytope. In particular, Eq. (8) holds since μ_T satisfies the global constraints.

Finally, we do not make any assumption here about the form of the model scores θ. Therefore, these results apply more generally, even when factor scores are obtained from non-linear functions of the inputs, such as deep neural networks (e.g., Chen et al., 2015).

4.1. When Relaxed Training is Better than Exact Training

Unfortunately, the bound in Eq. (10) might sometimes be loose. Indeed, to get the bound we have discarded the exact-hinge term (Eq. (8)), added the task-loss (Eq. (9)), and maximized the loss-augmented objective (Eq. (10)). At the same time, Eq. (7) provides a precise characterization of the integrality gap. Specifically, the gap is determined by the difference between the relaxed-hinge and the exact-hinge terms. This implies that even when the relaxed-hinge is not zero, a small integrality gap can still be attained if the exact-hinge is also large. In fact, the *only way* to get a large integrality gap is by setting the exact-hinge much smaller than the relaxed-hinge. But when can this happen? As we now show, it is less likely to happen when training with relaxed inference than when training with exact inference.

A key point is that the relaxed and exact hinge terms are upper bounded by the relaxed and exact *training objectives* respectively (the latter additionally depend on the task loss Δ). Therefore, minimizing the training objective will also likely reduce the corresponding hinge term (this is also demonstrated empirically in Section 7). Using this insight, we observe that relaxed training reduces the relaxed-hinge term without directly reducing the exact-hinge term, and thereby induces a small integrality gap. On the other hand, this also suggests that *exact training may actually increase the integrality gap*, since it reduces the exact-hinge without also reducing directly the relaxed-hinge term. This finding is consistent with previous empirical evidence. Specifically, Martins et al. (2009b, Table 2) showed that on

6. This is not true for random models (w), which often yield loose relaxations.

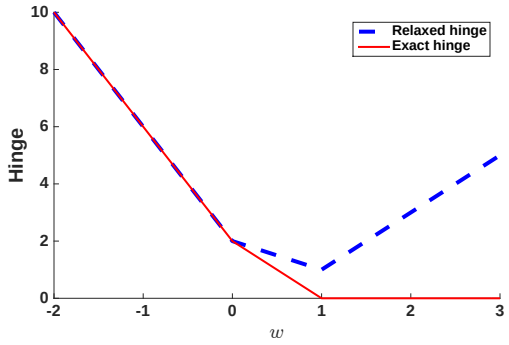

Figure 2: Exact- and relaxed- hinge (see Eq. (6)) as a function of w for the learning scenario in Section 4.1.

a dependency parsing problem, training with the relaxed objective achieved 92.9% integral solutions, while exact training achieved only 83.5% integral solutions. An even stronger effect was observed by Finley and Joachims (2008, Table 3) for multi-label classification, where relaxed training resulted in 99.6% integral instances, with exact training attaining only 17.7% ('Yeast' dataset).

In Section 7 we provide further empirical support for our explanation, however, we next show its possible limitation by providing a counter-example. The counter-example demonstrates that despite training with a relaxed objective, the exact-hinge can in some cases actually be *substantially smaller* than the relaxed-hinge, leading to a loose relaxation. Although this illustrates the limitations of the explanation above, we point out that the corresponding learning task is far from natural.

We construct a learning scenario where relaxed training obtains zero exact-hinge and non-zero relaxed-hinge, so the relaxation is not tight. Consider a model where $x \in \mathbb{R}^3$, $y \in \{0,1\}^3$, and the prediction is given by:

$$y(x;w) = \arg\max_y \Big(x_1 y_1 + x_2 y_2 + x_3 y_3 + w \left[\mathbb{1}\{y_1 \neq y_2\} + \mathbb{1}\{y_1 \neq y_3\} + \mathbb{1}\{y_2 \neq y_3\} \right] \Big). \tag{11}$$

The corresponding LP relaxation is then:

$$\begin{aligned} \max_{\mu \in \mathcal{M}_L} \Big(& x_1 \mu_1(1) + x_2 \mu_2(1) + x_3 \mu_3(1) \\ & + w \left[\mu_{12}(01) + \mu_{12}(10) + \mu_{13}(01) + \mu_{13}(10) + \mu_{23}(01) + \mu_{23}(10) \right] \Big). \end{aligned} \tag{12}$$

Next, we construct a training set where the first instance is: $x^{(1)} = (2,2,2), y^{(1)} = (1,1,0)$, and the second is: $x^{(2)} = (0,0,0), y^{(2)} = (1,1,0)$. Fig. 2 shows the relaxed and exact losses as a function of w, obtained by plugging Eq. (11) and Eq. (12) in Eq. (3) and Eq. (4), respectively.[7] Observe that $w = 1$ minimizes the relaxed objective (Eq. (4)). However, with this weight vector the relaxed-hinge for the second instance $(x^{(2)}, y^{(2)})$ is equal to 1, while the exact-hinge for this instance is 0 (the data is separable with $w = 1$). Consequently, there is an integrality gap of 1 for the second instance, and the relaxation is loose (the first

7. For simplicity we use $\ell = 0$ in this example, but a similar result holds with $\ell = 1/2 \cdot \ell_{\text{Hamming}}$.

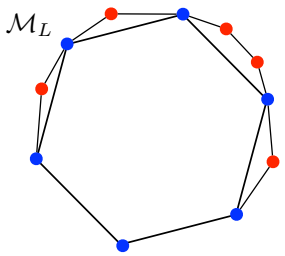

Figure 3: Illustration of the local marginal polytope $\mathcal{M}_L$, with its vertices partitioned into integral vertices $\mathcal{V}_I$ (blue), and fractional vertices $\mathcal{V}_F$ (red).

instance is actually tight). Notice that in this data the same output corresponds to two very different inputs.

5. Generalization of Tightness

Our argument in Section 4 concerns only the tightness of train instances. However, the empirical evidence discussed above pertains to test data. To bridge this gap, in this section we prove a generalization bound which shows that train tightness implies test tightness.

We first define a loss function which measures the lack of integrality (or, fractionality) of the LP solution for a given instance. To this end, we consider the discrete set of *vertices* of the local polytope $\mathcal{M}_L$ (excluding its convex hull), denoting by $\mathcal{V}_I$ and $\mathcal{V}_F$ the sets of fully-integral and non-integral (i.e., fractional) vertices,[8] respectively (so $\mathcal{V}_I \cap \mathcal{V}_F = \emptyset$, and $\mathcal{V}_I \cup \mathcal{V}_F$ consists of all vertices of $\mathcal{M}_L$);[9] see Fig. 3. Considering only vertices does not reduce generality, since the solution to a linear program is always at a vertex.

Next, let

$$I^*(w,x) \triangleq \max_{\mu \in \mathcal{V}_I} \theta(x;w)^\top \mu \qquad \text{and} \qquad F^*(w,x) \triangleq \max_{\mu \in \mathcal{V}_F} \theta(x;w)^\top \mu \tag{13}$$

denote the respective best integral and fractional scores. By convention, we set $F^*(w,x) \triangleq -\infty$ whenever $\mathcal{V}_F = \emptyset$. The fractionality of inference with (w,x) can be measured by the quantity

$$D(w,x) \triangleq F^*(w,x) - I^*(w,x). \tag{14}$$

Observe that $D(w,x) > 0$ whenever the LP has a fractional solution that is better than the integral solution. We can now define the *integrality loss*,

$$\mathcal{L}_0(w,x) \triangleq \begin{cases} 1 & D(w,x) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{15}$$

8. It is enough that one coordinate is fractional to belong to $\mathcal{V}_F$.
9. We assume that all feasible integral solutions are vertices of $\mathcal{M}_L$, which is the case for the type of relaxations considered here (see Wainwright and Jordan, 2008).

This loss function equals 1 if and only if the optimal fractional solution has a (strictly) higher score than the optimal integral solution. The loss will be 0 whenever the non-integral and integral optima are equal—that is, for our purpose we consider the relaxation to be tight in this case. The expected integrality loss measures the probability of obtaining a fractional LP solution (over draws of an input, x). Note that this loss ignores the ground truth assignment.

To support our generalization analysis, we define a related loss function, which we call the *integrality ramp loss*. For a predetermined margin parameter, γ, the integrality ramp loss is given by

$$\mathcal{L}_\gamma(w,x) \triangleq \begin{cases} 1 & D(w,x) > 0 \\ 1 + D(w,x)/\gamma & -\gamma < D(w,x) \leq 0 \\ 0 & D(w,x) \leq -\gamma \end{cases} . \tag{16}$$

Importantly, the integrality ramp loss upper-bounds the integrality loss. For the ramp loss to be zero, the best integral solution has to be better than the best fractional solution by a margin of at least γ, which is a stronger requirement than mere tightness. In Appendix A we give examples of models that are guaranteed to satisfy this requirement, and in Section 7 we also show this often happens in practice.

We point out that both $\mathcal{L}_0(w,x)$ and $\mathcal{L}_\gamma(w,x)$ are generally hard to compute, a point which we address in Section 6. For the time being, we are only interested in proving that tightness is a generalizing property, so we will not worry about computational efficiency.

We are now ready to state the main theorem of this section, a generalization bound for tightness. Our proof (deferred to Appendix B.1) uses a PAC-Bayesian analysis, similar to London et al. (2016), though the main result is stated for a deterministic predictor.

Theorem 1. *Let $\mathbb{D}$ denote a distribution over $\mathcal{X}$. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ denote a feature mapping such that $\sup_{x,y} \|\phi(x,y)\|_2 \leq B < \infty$. Then, for any $\gamma > 0$, $\delta \in (0,1)$ and $m \geq 1$, with probability at least $1-\delta$ over draws of $(x^{(1)}, \ldots, x^{(m)}) \in \mathcal{X}^m$, according to $\mathbb{D}^m$, every weight vector, w, with $\|w\|_2 \leq R < \infty$, satisfies*

$$\mathop{\mathbb{E}}_{x\sim\mathbb{D}}[\mathcal{L}_0(w,x)] \leq \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}_\gamma(w,x^{(i)}) + \frac{8}{m} + 2\sqrt{\frac{d\ln(mBR/\gamma) + \ln\frac{2}{\delta}}{2m}}. \tag{17}$$

Theorem 1 shows that if we observe high integrality (equivalently, low fractionality) on a finite sample of training data, then it is likely that integrality of test data will not be much lower, provided sufficient number of samples. It is worth noting that, though we focus on linear models to simplify our presentation, it is possible to extend this result to accommodate non-linear models (e.g., scores θ computed by a deep neural network), provided some assumptions are made on the smoothness of the model and loss.

As the following corollary states, Theorem 1 actually applies more generally to any two disjoint sets of vertices, and is not limited to $\mathcal{V}_I$ and $\mathcal{V}_F$.

Corollary 1. *Let $\mathcal{V}_\alpha$ be any set of vertices of $\mathcal{M}_L$ with at most α fractional values (where $0 \leq \alpha \leq 1$), and let $\bar{\mathcal{V}}_\alpha$ be the rest of the vertices of $\mathcal{M}_L$. Then Theorem 1 holds with $\mathcal{V}_\alpha$ and $\bar{\mathcal{V}}_\alpha$ replacing $\mathcal{V}_I$ and $\mathcal{V}_F$ in the definition of I^* and F^* in Eq. (13), respectively.*

For example, we can set $\mathcal{V}_\alpha$ to be any set of vertices with at most 10% fractional values, and $\bar{\mathcal{V}}_\alpha$ to be the rest of the vertices of $\mathcal{M}_L$. This gives a different meaning to the integrality loss, but the rest of our analysis holds unchanged. Consequently, our generalization result implies that it is likely to observe a similar portion of instances with at most 10% fractional values at test time as we did at training.

Moreover, Theorem 1 also holds in the presence of global constraints (e.g., spanning tree constraints). As mentioned in Section 4, the polytope $\mathcal{M}_L$ is replaced by its intersection with the global constraints polytope, but the rest of our derivation remains unchanged.

Note that the loss function in Eq. (15) does not measure the actual number of fractional values, nor their *distance* to integrality. In Section 6, we analyze a notion of tightness that accounts for the L1 distance to integrality.

Compared to the generalization bound of Kulesza and Pereira (2007), our bounds only consider the tightness of a prediction, ignoring label errors. Thus, for example, if learning happens to settle on a set of parameters in a tractable regime (e.g., supermodular potentials or stable instances (Makarychev et al., 2014)) for which the LP relaxation is tight for most training instances, our generalization bound guarantees that with high probability the LP relaxation will also be tight on most test instances. In contrast, in Kulesza and Pereira (2007), tightness on test instances can only be guaranteed when the training data is algorithmically separable (i.e., LP-relaxed inference predicts perfectly).

5.1. γ-Tight Relaxations

In this section we study the stronger notion of tightness required by our surrogate fractionality loss (Eq. (16)), and show examples of models that satisfy it.

Definition 1. An LP relaxation is called γ-*tight* if $I^*(w,x) \geq F^*(w,x) + \gamma$ (so $\mathcal{L}_\gamma(w,x) = 0$). That is, the best integral value is larger than the best non-integral value by at least γ.[10]

We focus on binary pairwise models and show two cases where the model is guaranteed to be γ-tight. Proofs are provided in Appendix A. Our first example involves *balanced* models, which are binary pairwise models that have supermodular scores, or can be made supermodular by "flipping" a subset of the variables (for more details, see Appendix A).

Proposition 1. *A balanced model with a unique optimum is $(\alpha/2)$-tight, where α is the difference between the best and second-best (integral) solutions.*

This result is of particular interest when learning structured predictors where the edge scores depend on the input. Whereas one could learn supermodular models by enforcing linear inequalities (Taskar et al., 2004), we know of no tractable means of ensuring the model is balanced. Instead, one could learn over the full space of models using LP relaxation. If the learned models are balanced on the training data, Proposition 1 together with Theorem 1 tell us that the LP relaxation is likely to be tight on test data as well.

Our second example regards models with singleton scores that are much stronger than the pairwise scores. Consider a binary pairwise model[11] in minimal representation, where $\bar{\theta}_i$ are node scores and $\bar{\theta}_{ij}$ are edge scores in this representation (see Appendix A for full

10. Notice that scaling up $\theta(w,x)$ will also increase γ, but our bound in Eq. (17) also grows with the norm of $\theta(w,x)$ (via the term BR). Therefore, we assume here that $\|\theta(w,x)\|_2$ is bounded.
11. This case easily generalizes to variables with more than 2 possible values.

details). Further, for each variable i, define the set of neighbors with *attractive* edges $N_i^+ = \{j \in N_i | \bar{\theta}_{ij} > 0\}$, and the set of neighbors with *repulsive* edges $N_i^- = \{j \in N_i | \bar{\theta}_{ij} < 0\}$.

Proposition 2. *If all variables satisfy the condition:*

$$\bar{\theta}_i \geq -\sum\nolimits_{j \in N_i^-} \bar{\theta}_{ij} + \beta, \quad \text{or} \quad \bar{\theta}_i \leq -\sum\nolimits_{j \in N_i^+} \bar{\theta}_{ij} - \beta$$

for some $\beta > 0$, then the model is $(\beta/2)$-tight.

Finally, we point out that in both of the examples above, the conditions can be verified efficiently and if they hold, the value of γ can be computed efficiently.

6. Analysis of the Integrality Distance

For structured prediction, the maximizer, μ_L, is often more important than the maximum value, $\theta^\top \mu_L$. That is, we do not really care whether the optimum of the relaxed problem equals that of the integral one; we just want relaxed inference to yield the optimal integral assignment—or, lacking that, an assignment that is "close to" the optimal integral one. If we assume that the relaxed problem has a unique solution, then an integrality gap of zero implies that the assignments are the same. However, lacking this assumption, there may be multiple, disparate solutions, so the assignments may differ. In general, it is difficult to characterize the distance between relaxed and exact assignments as a function of a nonzero integrality gap.

Thus, in addition to studying the integrality gap, we are also interested in what we will call the *integrality distance*, defined as $\|\mu_L - \mu_I\|_1$, which is the Manhattan distance between the maximizers of the relaxed and exact programs. The integrality distance is conceptually similar to *persistence* (see Wainwright and Jordan, 2008 for definition) in that a persistent fractional solution will have a subset of variables with zero integrality distance. The integrality distance is also related to the integrality gap, although the distance could sometimes be more useful: when the integrality distance is small, the relaxed solution is close to the exact solution, which is what we may ultimately care about; moreover, when the integrality distance is zero, the integrality gap must also be zero, regardless of whether we assume uniqueness.

In this section, we relate the integrality distance to several loss functions that are commonly analyzed in the literature on structured prediction. We then show that, similar to the integrality gap, the integrality distance also generalizes from an empirical sample to the population average. Moreover, we show that the integrality distance is upper-bounded by a constant multiple of the structured hinge loss—a convex loss function that is commonly used for training. Importantly, unlike the bound in Theorem 1, this bound is computationally tractable. Combining these results, we obtain a high-probability bound on the expected integrality distance that can be efficiently evaluated from training data, and whose additive error decreases with the number of examples. Finally, a simple argument shows how this bound applies to an integral rounding of a fractional solution.

6.1. Structured Loss Functions

We will focus on the integrality distance of the singleton (i.e., node) marginals, denoted $\mu_{\mathrm{u}} \triangleq (\mu_i)_{i=1}^n$, since they are sufficient for decoding a labeling, y. Let

$$D_1(\mu, \mu') \triangleq \frac{1}{2n} \left\| \mu_{\mathrm{u}} - \mu'_{\mathrm{u}} \right\|_1 \tag{18}$$

denote the normalized Manhattan distance. When both inputs are integral, D_1 is equivalent to the normalized Hamming distance.

Given a model, w, an input, x, and an assignment, μ, let

$$\mathcal{L}_1(w, x, \mu) \triangleq D_1\left(\mu, \mu_L(x; w)\right) \tag{19}$$

denote the L_1 *loss*. This loss function is a generalization of the Hamming loss, which is commonly used to measure the prediction error of exact inference. If the third argument is a reference (i.e., "ground truth") labeling, μ_T, then $\mathcal{L}_1(w, x, \mu_T)$ measures the prediction error of approximate inference. However, if the third argument is the exact, integral MAP state, μ_I, then $\mathcal{L}_1(w, x, \mu_I)$ is the normalized integrality distance. This latter quantity is what we will focus on upper-bounding.

Let

$$\mathcal{L}_{\mathrm{h}}(w, x, \mu) \triangleq \max_{\mu' \in \mathcal{M}_L} D_1(\mu, \mu') + \theta^\top \left(\mu' - \mu\right). \tag{20}$$

denote a loss function commonly referred to as the (relaxed) *structured hinge loss*. This loss is minimized when μ scores higher than all alternate assignments, μ', by a margin that is at least $D_1(\mu, \mu')$. Note that when μ is the exact MAP state, the structured hinge loss computes a *loss-augmented* integrality gap, using Manhattan distance for loss augmentation (see Eq. (4)).

A related loss function is the (relaxed) *structured ramp loss*,

$$\mathcal{L}_{\mathrm{r}}(w, x, \mu) \triangleq \max_{\mu' \in \mathcal{M}_L} D_1(\mu, \mu') + \theta^\top \left(\mu' - \mu_L(x; w)\right), \tag{21}$$

which can be considered a normalized version of the hinge loss. $\mathcal{L}_{\mathrm{r}}$ is bounded by $[0, 1]$, whereas $\mathcal{L}_{\mathrm{h}}$ might be unbounded (depending on the features and weights).

The hinge loss is often used in max-margin training, since it is convex in w. The ramp loss is not convex in w, but it is bounded, Lipschitz, and has a convenient relationship to the L_1 (or Hamming) and hinge losses:

$$\mathcal{L}_1(w, x, \mu) \leq \mathcal{L}_{\mathrm{r}}(w, x, \mu) \leq \mathcal{L}_{\mathrm{h}}(w, x, \mu). \tag{22}$$

Thus, the ramp loss is often used as an analytical tool to derive generalization bounds, such as those that follow.

6.2. Generalization Bound

Similar to Section 5, we now show that the integrality distance on a training sample generalizes to the data distribution. Like Theorem 1, the proof (in Appendix B.2) uses PAC-Bayesian analysis.

Theorem 2. *Let $\mathbb{D}$ denote a distribution over $\mathcal{X}$. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ denote a feature mapping such that $\sup_{x,y} \|\phi(x,y)\|_2 \leq B < \infty$; and μ_I is defined in Eq. (5). Then, for any $\delta \in (0,1)$ and $m \geq 1$, with probability at least $1-\delta$ over draws of $(x^{(1)}, \ldots, x^{(m)}) \in \mathcal{X}^m$, according to $\mathbb{D}^m$, every weight vector, w, with $\|w\|_2 \leq R < \infty$, satisfies*

$$\mathop{\mathbb{E}}_{x \sim \mathbb{D}}[\mathcal{L}_1(w, x, \mu_I)] \leq \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_r(w, x^{(i)}, \mu_I^{(i)}) + \frac{8}{m} + 2\sqrt{\frac{d \ln(mBR) + \ln \frac{2}{\delta}}{2m}}. \tag{23}$$

Further, Eq. (23) holds when $\mathcal{L}_r$ is replaced with $\mathcal{L}_h$.

Theorem 2 says that the integrality distance on the training set generalizes to future examples. More precisely, the expected integrality distance on a random instance is upper-bounded by the average integrality ramp (or hinge) loss on the training set, plus two terms that vanish as the number of training examples grows. Thus, the more training data we have, the better we can estimate the expected integrality distance.

Remark 1. There is nothing special about μ_I to Theorem 2. Indeed, we could use any integral assignment as a reference labeling for the loss functions and the proof would be the same. For example, we could replace μ_I with μ_T (a ground truth labeling) and obtain a risk bound for learning with approximate inference, which is a well-studied topic (e.g., Kulesza and Pereira, 2007; London et al., 2016).

6.3. Relationship to Max-Margin Training

In practice, computing the integrality loss is generally infeasible, since it requires exact inference. Therefore, the upper bounds in Theorems 1 and 2 cannot be evaluated. However, the empirical relaxed hinge loss with respect to the ground truth labels *can* be evaluated efficiently. In this section, we show how minimizing this quantity actually minimizes the integrality distance. That is, max-margin training with approximate inference—which is commonly used anyway to learn graphical models—reduces not only the prediction error, but also the inference approximation error.

The key insight that enables this result comes from the following technical lemma.

Lemma 1. *For any w and x, if μ_T is the reference (ground truth) labeling of x and μ_I is the exact MAP state under w, then*

$$\mathcal{L}_h(w, x, \mu_I) \;\leq\; 2\,\mathcal{L}_h(w, x, \mu_T), \tag{24}$$

meaning the integrality hinge loss is at most twice the hinge loss with respect to the true labeling.

Proof First, we decompose the integrality hinge loss as follows:

$$\begin{aligned}
\mathcal{L}_{\mathrm{h}}(w, x, \mu_I) &= \max_{\mu \in \mathcal{M}_L} D_1(\mu_I, \mu) + \theta^\top (\mu - \mu_I) \\
&\leq \max_{\mu \in \mathcal{M}_L} D_1(\mu_I, \mu) + \theta^\top (\mu - \mu_T) \\
&\leq D_1(\mu_I, \mu_T) + \max_{\mu \in \mathcal{M}_L} D_1(\mu_T, \mu) + \theta^\top (\mu - \mu_T) \\
&= D_1(\mu_I, \mu_T) + \mathcal{L}_{\mathrm{h}}(w, x, \mu_T).
\end{aligned} \tag{25}$$

The second term on the right-hand side is the hinge loss of the approximate predictor with respect to the true labeling, which can be evaluated efficiently. The first term on the right-hand side is the Hamming loss of exact inference, which cannot be evaluated efficiently. However, this latter quantity can be upper-bounded as follows:

$$\begin{aligned} D_1(\mu_I, \mu_T) &\leq \max_{\mu \in \mathcal{M}} D_1(\mu, \mu_T) + \theta^\top (\mu - \mu_T) \\ &\leq \max_{\mu \in \mathcal{M}_L} D_1(\mu, \mu_T) + \theta^\top (\mu - \mu_T) \\ &= \mathcal{L}_{\mathrm{h}}(w, x, \mu_T). \end{aligned} \tag{26}$$

Combining Eq. (25) and (26) completes the proof. ■

Note that Lemma 1 also yields an upper bound on the integrality ramp loss, since it is upper-hounded by the integrality hinge loss.

Using Lemma 1, we thus obtain the following corollary of Theorem 2.

Corollary 2. *Let $\mathbb{D}$ denote a distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ denote a feature mapping such that $\sup_{x,y} \|\phi(x, y)\|_2 \leq B < \infty$. Then, for any $\delta \in (0, 1)$ and $m \geq 1$, with probability at least $1 - \delta$ over draws of $(x^{(1)}, y^{(1)}, \ldots, x^{(m)}, y^{(m)}) \in (\mathcal{X} \times \mathcal{Y})^m$, according to $\mathbb{D}^m$, every weight vector, w, with $\|w\|_2 \leq R < \infty$, satisfies*

$$\mathop{\mathbb{E}}_{x \sim \mathbb{D}}[\mathcal{L}_1(w, x, \mu_I)] \leq \frac{2}{m} \sum_{i=1}^{m} \mathcal{L}_h(w, x^{(i)}, \mu_T^{(i)}) + \frac{8}{m} + 2\sqrt{\frac{d \ln(mBR) + \ln \frac{2}{\delta}}{2m}}, \tag{27}$$

where $\mu_T^{(i)}$ denotes the integral vector corresponding to the ground-truth labeling $y^{(i)}$.

Corollary 2 says that max-margin training with relaxed inference directly minimizes the integrality distance on future examples. Importantly, if the constants B and R are known, then this bound can be efficiently evaluated from training data. It is worth noting that Corollary 2 actually holds for *any* integral assignment, not just the ground truth labels. Nonetheless, we feel the bound is more insightful when stated with respect to the ground truth labels, which are given in the learning setup. In this case $\mathbb{D}$ is defined as a joint distribution over $\mathcal{X}$ and $\mathcal{Y}$, and the bound holds with high probability over draws of both inputs and labels. It is also worth noting that Corollary 2 holds in the presence of global constraints—provided the ground truth (or whichever integral assignments are used) are in the feasible set.

6.4. Decoding a Solution

When the solution to an LP relaxation is fractional, we often round the solution to an integral assignment. Rounding schemes have been studied extensively (e.g., Raghavan and Tompson, 1987; Kleinberg and Tardos, 2002; Chekuri et al., 2004; Ravikumar et al., 2010). Arguably, the simplest method is to select the local assignments $\mu_i(y_i)$ with the highest values. One question that arises is how far the rounding, denoted $\mu_R(x; w)$, is from the exact solution; once this relationship is determined, one can apply our prior generalization analysis to the rounding. It turns out that the distance from μ_R to μ_I can be upper-bounded by a multiple of the integrality distance.

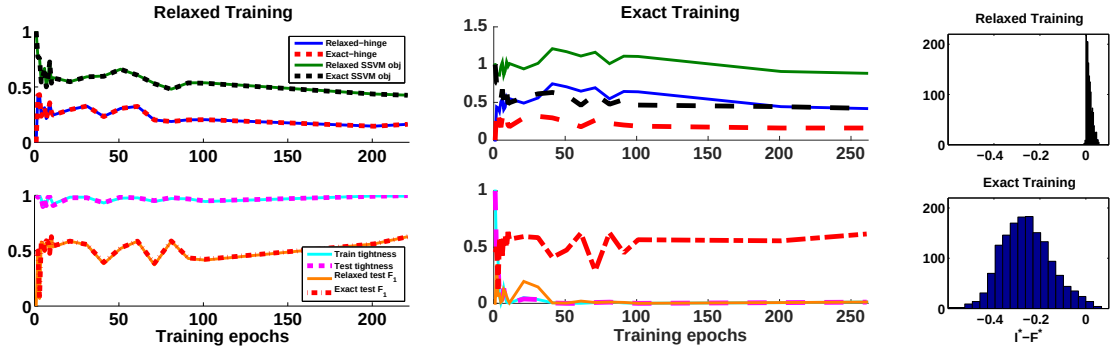

Figure 4: Training with the 'Yeast' multi-label classification dataset. Various quantities of interest are shown as a function of training iterations. (Left) Training with LP relaxation. (Middle) Training with ILP. (Right) Integrality margin (bin widths are scaled differently).

Lemma 2. *Suppose that every output variable has the same domain—i.e.,* $\mathcal{Y}_1 = \mathcal{Y}_2 = \ldots = \mathcal{Y}_n$*—and that each domain has size* k*. If* $\mu_R(x; w)$ *is the rounding of the fractional solution,* $\mu_L(x; w)$*, then*

$$D_1(\mu_R, \mu_I) \;\leq\; k\, D_1(\mu_L, \mu_I). \tag{28}$$

Proof Consider any output variable. If the fractional solution assigns the majority of the local belief to the "correct" label (i.e., the label chosen by exact inference), then the rounding of that variable will be exact. However, if the fractional solution puts most of the local belief on an "incorrect" label, then the rounding of that variable will have D_1 distance 1 from the correct label. Since the incorrect label must have had a fractional value of at least $1/k$, it follows that the fractional solution has D_1 distance at least $1/k$, which is no less than $1/k$ that of the rounding. Applying this logic to every variable completes the proof. ■

Lemma 2 can be combined with Corollary 2 to generate bounds on the expected integrality distance of rounding; the bound simply scales by k.

7. Experiments

In this section we present some numerical results to support our theoretical analysis. We run experiments for a multi-label classification task and an image segmentation task. For training we have implemented the block-coordinate Frank-Wolfe algorithm for structured SVM (Lacoste-Julien et al., 2013), using GLPK as the LP solver. We use a standard L_2 regularizer, chosen via cross-validation.

7.1. Multi-Label Classification

For multi-label classification we adopt the experimental setting of Finley and Joachims (2008). In this setting labels are represented by binary variables, the model consists of

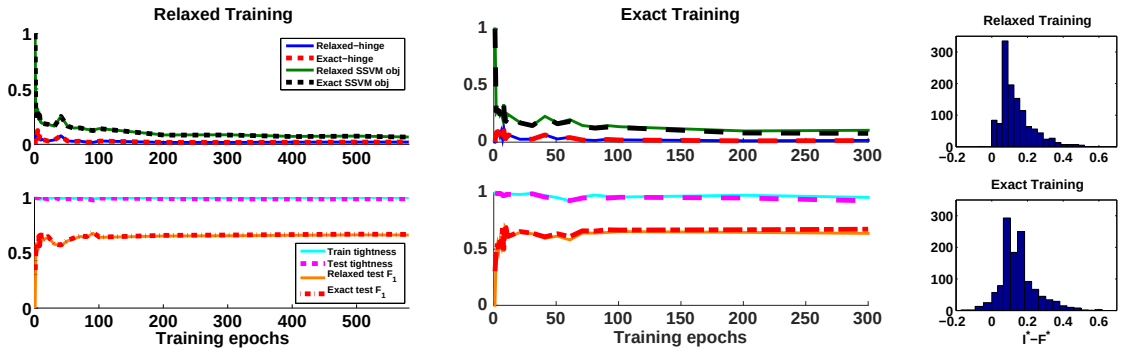

Figure 5: Training with the 'Scene' multi-label classification dataset. Various quantities of interest are shown as a function of training iterations. (Left) Training with LP relaxation. (Middle) Training with ILP. (Right) Integrality margin.

singleton and pairwise factors forming a fully connected graph over the labels, and the task loss is the normalized Hamming distance.

Fig. 4 shows relaxed and exact training iterations for the 'Yeast' dataset (14 labels). We plot the relaxed and exact hinge terms (Eq. (6)), the exact and relaxed SSVM training objectives[12] (Eq. (3) and Eq. (4), respectively), fraction of train and test instances having integral solutions, as well as test accuracy (measured by F_1 score). We use a simple scheme to round fractional solutions found with relaxed inference. First, we note that the relaxed-hinge values are nicely correlated with the relaxed training objective, and likewise the exact-hinge is correlated with the exact objective (left and middle, top). Second, observe that with relaxed training, the relaxed-hinge and the exact-hinge are very close (left, top), so the integrality gap, given by their difference, remains small (almost 0 here). On the other hand, with exact training there is a large integrality gap (middle, top). Indeed, we can see that the percentage of integral solutions is almost 100% for relaxed training (left, bottom), and close to 0% with exact training (middle, bottom). In Fig. 4 (right) we also show histograms of the difference between the optimal integral and fractional values, i.e., the integrality margin $(I^*(w,x) - F^*(w,x))$, under the final learned w for all training instances. It can be seen that with relaxed training this margin is positive (although small), while exact training results in larger negative values. Finally, we note that train and test integrality levels are very close to each other, almost indistinguishable (left and middle, bottom), which provides empirical support to our generalization result from Section 5.

We next train a model using random labels (with similar label counts as the true data). In this setting the learned model obtains 100% tight training instances (not shown), which supports our observation that any integral point can be used in place of the ground-truth, and that accuracy is not important for tightness. Finally, in order to verify that tightness is not coincidental, we test the tightness of the relaxation induced by a random weight vector w. We find that random models are never tight (in 20 trials), which shows that tightness of the relaxation does not come by chance.

12. The displayed objective values are averaged over train instances and exclude regularization.

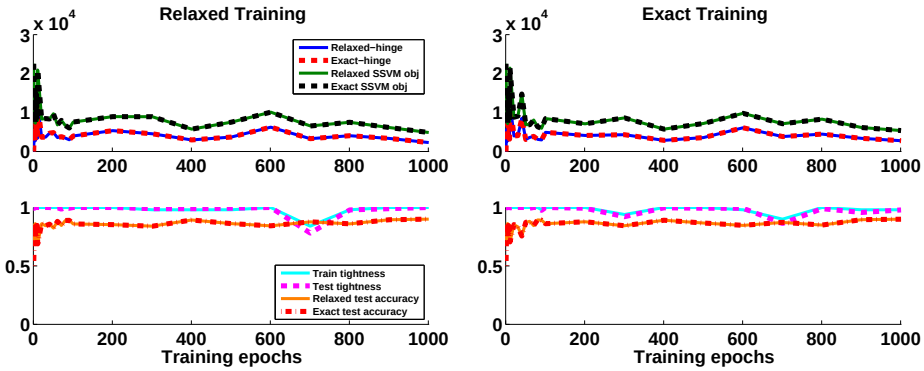

Figure 6: Training for foreground-background segmentation with the Weizmann Horse dataset. Various quantities of interest are shown as a function of training iterations. (Left) Training with LP relaxation. (Right) Training with ILP.

We now proceed to perform experiments on the 'Scene' dataset (6 labels). The results, in Fig. 5, differ from the 'Yeast' results in case of exact training (middle). Specifically, we observe that in this case the relaxed-hinge and exact-hinge are close in value (middle, top), as for relaxed training (left, top). As a consequence, the integrality gap is very small and the relaxation is tight for almost all train (and test) instances. These results illustrate that sometimes optimizing the exact objective can reduce the relaxed objective (and relaxed-hinge) as well. Further, in this setting we observe a larger integrality margin (right), namely the integral optimum is strictly better than the fractional one.

We conjecture that the LP instances are easy in this case due to the dominance of the singleton scores.[13] Specifically, the features provide a strong signal which allows label assignment to be decided mostly based on the local score, with little influence coming from the pairwise terms. To test this conjecture we inject Gaussian noise into the input features, forcing the model to rely more on the pairwise interactions. We find that with the noisy singleton scores the results are indeed more similar to the 'Yeast' dataset, where a large integrality gap is observed and fewer instances are tight (see Appendix C in the supplement).

7.2. Image Segmentation

Finally, we conduct experiments on a foreground-background segmentation problem using the Weizmann Horse dataset (Borenstein et al., 2004). The data consists of 328 images, of which we use the first 50 for training and the rest for testing. Here a binary output variable is assigned to each pixel, and there are $\sim 58K$ variables per image on average. We extract singleton and pairwise features as described in Domke (2013). Fig. 6 shows the same quantities as in the multi-label setting, except for the accuracy measure—here we compute the percentage of correctly classified pixels rather than F_1. We observe a very similar behavior to that of the 'Scene' multi-label dataset (Fig. 5). Specifically, both relaxed and exact training produce a small integrality gap and high percentage of tight instances.

13. With ILP training, the condition in Proposition 2 is satisfied for 65% of all variables, although only 1% of the training instances satisfy it for all their variables.

Unlike the 'Scene' dataset, here only 1.2% of variables satisfy the condition in Proposition 2 (using LP training). In all of our experiments the learned model scores were never balanced (Proposition 1), although for the segmentation problem we believe the models learned are close to balanced, both for relaxed and exact training.

8. Conclusion

In this paper we present a theoretical analysis of the tightness of LP relaxations often observed in structured prediction applications. Our analysis is based on a careful examination of the integrality gap, the integrality distance, and their relation to the training objective. It shows how training with LP relaxations, although designed with accuracy considerations in mind, also induces tightness of the relaxation. Our derivation also suggests that exact training may sometimes have the opposite effect, increasing the integrality gap.

To explain tightness at test time, we show that tightness generalizes in the following two senses: first, if most training predictions are integral, then most test instances will also be integral; secondly, if training predictions are on average "close to" integral, then test predictions will be similarly close to integral in expectation.

Acknowledgments

AW acknowledges support from the David MacKay Newton research fellowship at Darwin College, The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074, and the Leverhulme Trust via the CFI. DS acknowledges support from NSF CCF-1723344 and NSF CAREER award #1350965.

Appendix A. γ-Tight LP Relaxations

In this section we provide full derivations for the results in Section 5.1. We make extensive use of the results in Weller et al. (2016), some of which are restated here for completeness. We start by defining a model in minimal representation, which will be convenient for the derivations that follow. Specifically, in the case of binary variables ($y_i \in \{0,1\}$) with pairwise factors, we define a value η_i for each variable, and a value η_{ij} for each pair. The mapping between the over-complete vector μ and the minimal vector η is as follows. For singleton factors, we have:

$$\mu_i = \begin{pmatrix} 1-\eta_i \\ \eta_i \end{pmatrix}$$

Similarly, for the pairwise factors, we have:

$$\mu_{ij} = \begin{pmatrix} 1+\eta_{ij}-\eta_i-\eta_j & \eta_j-\eta_{ij} \\ \eta_i-\eta_{ij} & \eta_{ij} \end{pmatrix},$$

The corresponding mapping to minimal parameters is then:

$$\bar{\theta}_i = \theta_i(1) - \theta_i(0) + \sum_{j \in N_i} (\theta_{ij}(1,0) - \theta_{ij}(0,0))$$
$$\bar{\theta}_{ij} = \theta_{ij}(1,1) + \theta_{ij}(0,0) - \theta_{ij}(0,1) - \theta_{ij}(1,0)$$

In this representation, the LP relaxation is given by (up to constants):

$$\max_{\eta \in \mathbb{L}} f(\eta) := \sum_{i=1}^{n} \bar{\theta}_i \eta_i + \sum_{ij \in \mathcal{E}} \bar{\theta}_{ij} \eta_{ij}$$

where $\mathbb{L}$ is the appropriate transformation of $\mathcal{M}_L$ to the equivalent reduced space of η:

$$\begin{aligned} 0 &\leq \eta_i \leq 1 && \forall i \\ \max(0, \eta_i + \eta_j - 1) &\leq \eta_{ij} \leq \min(\eta_i, \eta_j) && \forall ij \in \mathcal{E} \end{aligned}$$

If $\bar{\theta}_{ij} > 0$ ($\bar{\theta}_{ij} < 0$), then the edge is called *attractive* (*repulsive*). If all edges are attractive, then the LP relaxation is known to be tight (Wainwright and Jordan, 2008). When not all edges are attractive, in some cases it is possible to make them attractive by *flipping* a subset of the variables ($y_i \leftarrow 1 - y_i$), which flips the signs of edge potentials for edges with exactly one end in the flipped subset.[14] In such cases the model is called *balanced*.

In the sequel we will make use of the known fact that all vertices of the local polytope are half-integral (take values in $\{0, \frac{1}{2}, 1\}$) (Padberg, 1989). We are now ready to prove the propositions (restated here for convenience).

14. The flip-set, if exists, is easy to find by making a single pass over the graph (see Weller (2015) for more details).

A.1. Proof of Proposition 1

Proposition 1 *A balanced model with a unique optimum is $(\alpha/2)$-tight, where α is the difference between the best and second-best (integral) solutions.*

Proof Weller et al. (2016) define for a given variable i the function $F^i_{\mathbb{L}}(z)$, which returns for every $0 \leq z \leq 1$ the constrained optimum:

$$F^i_{\mathbb{L}}(z) = \max_{\substack{\eta \in \mathbb{L} \\ \eta_i = z}} f(\eta)$$

Given this definition, they show that for a balanced model, $F^i_{\mathbb{L}}(z)$ is a *linear function* (Weller et al., 2016, Theorem 6).

Let m be the optimal score, let η^1 be the unique optimum integral vertex in minimal form so $f(\eta^1) = m$, and by assumption any other integral vertex has value at most $m - \alpha$. Denote the state of η^1 at coordinate i by $z^* = \eta^1_i$, and consider computing the constrained optimum holding η_i to various states. By assumption, any other integral vertex has value at most $m - \alpha$, therefore,

$$F^i_{\mathbb{L}}(z^*) = m$$
$$F^i_{\mathbb{L}}(1 - z^*) \leq m - \alpha$$

(the second line holds with equality if there exists a second-best solution η^2 s.t. $\eta^2_i \neq \eta^1_i$). Since $F^i_{\mathbb{L}}(z)$ is a linear function, we have that:

$$F^i_{\mathbb{L}}(1/2) \leq m - \alpha/2 \tag{29}$$

Next, towards contradiction, suppose that there exists a fractional vertex η^f with value $f(\eta^f) > m - \alpha/2$. Let j be a fractional coordinate, so $\eta^f_j = \frac{1}{2}$ (since vertices are half-integral). Our assumption implies that $F^j_{\mathbb{L}}(1/2) > m - \alpha/2$, but this contradicts Eq. (29). Therefore, we conclude that any fractional solution has value at most $f(\eta^f) \leq m - \alpha/2$. ■

It is possible to check in polynomial time if a model is balanced, if it has a unique optimum, and compute α. This can be done by computing the difference in value to the second-best. In order to find the second-best: one can constrain each variable in turn to differ from the state of the optimal solution, and recompute the MAP solution; finally, take the maximum over all these trials.

A.2. Proof of Proposition 2

Proposition 2 *If all variables satisfy the condition:*

$$\bar{\theta}_i \geq -\sum_{j \in N_i^-} \bar{\theta}_{ij} + \beta, \quad \text{or} \quad \bar{\theta}_i \leq -\sum_{j \in N_i^+} \bar{\theta}_{ij} - \beta$$

for some $\beta > 0$, then the model is $(\beta/2)$-tight.

Proof For any binary pairwise models, given singleton terms $\{\eta_i\}$, the *optimal* edge terms are given by (for details see Weller et al., 2016):

$$\eta_{ij}(\eta_i, \eta_j) = \begin{cases} \min(\eta_i, \eta_j) & \text{if } \bar{\theta}_{ij} > 0 \\ \max(0, \eta_i + \eta_j - 1) & \text{if } \bar{\theta}_{ij} < 0 \end{cases}$$

Now, consider a variable i and let N_i be the set of its neighbors in the graph. Further, define the sets $N_i^+ = \{j \in N_i | \bar{\theta}_{ij} > 0\}$ and $N_i^- = \{j \in N_i | \bar{\theta}_{ij} < 0\}$, corresponding to attractive and repulsive edges, respectively. We next focus on the parts of the objective affected by the value at η_i (recomputing optimal edge terms); recall that all vertices are half-integral:

$\eta_i = 1$	$\eta_i = 1/2$	$\eta_i = 0$
$\bar{\theta}_i + \sum_{\substack{j \in N_i^+ \\ \eta_j = 1}} \bar{\theta}_{ij} + \frac{1}{2} \sum_{\substack{j \in N_i^+ \\ \eta_j = \frac{1}{2}}} \bar{\theta}_{ij} + \sum_{\substack{j \in N_i^- \\ \eta_j = 1}} \bar{\theta}_{ij} + \frac{1}{2} \sum_{\substack{j \in N_i^- \\ \eta_j = \frac{1}{2}}} \bar{\theta}_{ij}$	$\frac{1}{2}\bar{\theta}_i + \frac{1}{2} \sum_{\substack{j \in N_i^+ \\ \eta_j \in \{\frac{1}{2}, 1\}}} \bar{\theta}_{ij} + \frac{1}{2} \sum_{\substack{j \in N_i^- \\ \eta_j = 1}} \bar{\theta}_{ij}$	0

It is easy to verify that the condition $\bar{\theta}_i \geq -\sum_{j \in N_i^-} \bar{\theta}_{ij} + \beta$ guarantees that $\eta_i = 1$ in the optimal solution. We next bound the difference in objective values resulting from setting $\eta_i = 1/2$.

$$\Delta f = \frac{1}{2} \left(\bar{\theta}_i + \sum_{\substack{j \in N_i^+ \\ \eta_j = 1}} \bar{\theta}_{ij} + \sum_{\substack{j \in N_i^- \\ \eta_j \in \{\frac{1}{2}, 1\}}} \bar{\theta}_{ij} \right) \geq \frac{1}{2} \left(\bar{\theta}_i + \sum_{j \in N_i^-} \bar{\theta}_{ij} \right) \geq \beta/2$$

Similarly, when $\bar{\theta}_i \leq -\sum_{j \in N_i^+} \bar{\theta}_{ij} - \beta$, then $\eta_i = 0$ in any optimal solution. The difference in objective values from setting $\eta_i = 1/2$ in this case is:

$$\Delta f = -\frac{1}{2} \left(\bar{\theta}_i + \sum_{\substack{j \in N_i^+ \\ \eta_j \in \{\frac{1}{2}, 1\}}} \bar{\theta}_{ij} + \sum_{\substack{j \in N_i^- \\ \eta_j = 1}} \bar{\theta}_{ij} \right) \geq -\frac{1}{2} \left(\bar{\theta}_i + \sum_{j \in N_i^+} \bar{\theta}_{ij} \right) \geq \beta/2$$

Notice that for more fractional coordinates the difference in values can only increase, so in any case the fractional solution is worse by at least $\beta/2$. ■

Appendix B. Generalization Bound Proofs

To prove Theorems 1 and 2, we will use the following PAC-Bayes bound. There are other ways of proving generalization bounds for structured predictors (e.g., Weiss and Taskar, 2010), and other PAC-Bayesian analyses of structured prediction (e.g., McAllester, 2007), but we prefer the following for its simplicity. Our analysis is based on London et al.'s (2016), with certain simplifications due to differences in our objective.

Lemma 3. *Let $\mathbb{D}$ denote a distribution over an instance space, $\mathcal{Z}$. Let $\mathcal{H}$ denote a hypothesis class. Let $\mathcal{L} : \mathcal{H} \times \mathcal{Z} \to [0, 1]$ denote a bounded loss function. Let $\mathbb{P}$ denote a fixed prior distribution over $\mathcal{H}$. Then, for any $\delta \in (0, 1)$ and $m \geq 1$, with probability at least $1 - \delta$ over draws of $(Z^{(1)}, \ldots, Z^{(m)}) \in \mathcal{Z}^m$, according to $\mathbb{D}^m$, every posterior distribution, $\mathbb{Q}$, over $\mathcal{H}$, satisfies*

$$\mathop{\mathbb{E}}_{Z \sim \mathbb{D}} \mathop{\mathbb{E}}_{h \sim \mathbb{Q}} [\mathcal{L}(h, Z)] \leq \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{h \sim \mathbb{Q}} [\mathcal{L}(h, Z^{(i)})] + 2\sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \ln \frac{2}{\delta}}{2m}} \tag{30}$$

Proof To simplify notation, let

$$\varphi(h, Z^{(i)}) \triangleq \frac{1}{m}\left(\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}[\mathcal{L}(h,Z)] - \mathcal{L}(h, Z^{(i)})\right).$$

For any free parameter, $\epsilon \in \mathbb{R}$, observe that

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\mathcal{L}(h,Z)] - \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\mathcal{L}(h,Z^{(i)})] = \frac{1}{\epsilon}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\sum_{i=1}^{m}\epsilon\,\varphi(h,Z^{(i)})\right]. \tag{31}$$

The next step uses Donsker and Varadhan's (1975) *change of measure* inequality, which states that, if X is a random variable taking values in Ω, then for any two distributions, $\mathbb{P}$ and $\mathbb{Q}$, on Ω,

$$\mathop{\mathbb{E}}_{X\sim\mathbb{Q}}[X] \le D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{X\sim\mathbb{P}}\left[e^{X}\right].$$

Applying change of measure to the righthand side of Equation 31, we have

$$\frac{1}{\epsilon}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\sum_{i=1}^{m}\epsilon\,\varphi(h,Z^{(i)})\right] \le \frac{1}{\epsilon}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[\exp\left(\sum_{i=1}^{m}\epsilon\,\varphi(h,Z^{(i)})\right)\right]\right). \tag{32}$$

By Markov's inequality, with probability $1-\delta$ over draws of a training set, $\mathbf{Z} \triangleq (Z^{(1)},\ldots,Z^{(m)})$, according to $\mathbb{D}^m$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[\exp\left(\sum_{i=1}^{m}\epsilon\,\varphi(h,Z^{(i)})\right)\right] &\le \frac{1}{\delta}\mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}^m}\mathop{\mathbb{E}}_{h\sim\mathbb{P}}\left[\exp\left(\sum_{i=1}^{m}\epsilon\,\varphi(h,Z^{(i)})\right)\right] \\
&= \frac{1}{\delta}\mathop{\mathbb{E}}_{h\sim\mathbb{P}}\mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathbb{D}^m}\left[\prod_{i=1}^{m}\exp\left(\epsilon\,\varphi(h,Z^{(i)})\right)\right] \\
&= \frac{1}{\delta}\mathop{\mathbb{E}}_{h\sim\mathbb{P}}\prod_{i=1}^{m}\mathop{\mathbb{E}}_{Z^{(i)}\sim\mathbb{D}}\left[\exp\left(\epsilon\,\varphi(h,Z^{(i)})\right)\right].
\end{aligned} \tag{33}$$

In the last line, we leveraged the fact that the expectation of a product of i.i.d. random variables (in this case, $\varphi(h, Z^{(i)})$) is the product of their expectations. To upper-bound each expectation, we use Hoeffding's inequality, which states that if X is a zero-mean random variable, such that $a \le X \le b$ almost surely, then, for all $\epsilon \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\epsilon X}\right] \le \exp\left(\frac{\epsilon^2(b-a)^2}{8}\right).$$

Since $\varphi(h, Z^{(i)})$ has mean zero, and

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}[\mathcal{L}(h,Z)] - \frac{1}{m} \le \varphi(h,Z^{(i)}) \le \mathop{\mathbb{E}}_{Z\sim\mathbb{D}}[\mathcal{L}(h,Z)] - 0,$$

we have that

$$\mathop{\mathbb{E}}_{Z^{(i)}\sim\mathbb{D}}\left[\exp\left(\epsilon\,\varphi(h,Z^{(i)})\right)\right] \le \exp\left(\frac{\epsilon^2}{8m^2}\right). \tag{34}$$

Combining Equations 31 to 34, we have for any $\epsilon \in \mathbb{R}$, with probability at least $1-\delta$, all $\mathbb{Q}$ satisfy

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\mathcal{L}(h,Z)] - \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\mathcal{L}(h,Z^{(i)})] \leq \frac{1}{\epsilon}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta}\right) + \frac{\epsilon}{8m}. \tag{35}$$

The ϵ that minimizes Equation 35 is straightforward: setting the derivative equal to 0 and solving for $\epsilon^\star$, we have $\epsilon^\star = \sqrt{8m(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta})}$. Unfortunately, this optimal value depends on the choice of posterior, $\mathbb{Q}$, which means that the bound would only hold for a single posterior, not *all* posteriors simultaneously. To derive a bound that holds for all posteriors, we employ a covering technique, attributed to Seldin et al. (2012). We first define a countably infinite sequence of ϵ values. For each value, we assign an exponentially decreasing probability that Equation 35 fails to hold, such that with probability at least $1-\delta$ it holds for all values simultaneously. Then, for any posterior, we select an index, $j^\star$, that approximately optimizes Equation 35, and use lower and upper bounds on $\epsilon_{j^\star}$ to simplify the bound.

For $j = 0, 1, 2, \ldots$, let

$$\epsilon_j \triangleq 2^j\sqrt{8m\ln\frac{2}{\delta}}. \tag{36}$$

For each ϵ_j, we assign $\delta_j \triangleq \delta 2^{-(j+1)}$ probability that Equation 35 does not hold, substituting (ϵ_j, δ_j) for (ϵ, δ). Thus, with probability at least $1 - \sum_{j=0}^{\infty}\delta_j = 1 - \delta\sum_{j=0}^{\infty}2^{-(j+1)} = 1-\delta$, all $j = 0, 1, 2, \ldots$ satisfy

$$\mathop{\mathbb{E}}_{Z\sim\mathbb{D}} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\mathcal{L}(h,Z)] - \frac{1}{m}\sum_{i=1}^{m} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\mathcal{L}(h,Z^{(i)})] \leq \frac{1}{\epsilon_j}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_j}\right) + \frac{\epsilon_j}{8m}$$

For any given posterior, $\mathbb{Q}$, we choose an index, $j^\star$, by taking

$$j^\star \triangleq \left\lfloor \frac{1}{2\ln 2}\ln\left(\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})}{\ln(2/\delta)} + 1\right)\right\rfloor,$$

which implies

$$\sqrt{2m\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right)} \leq \epsilon_{j^\star} \leq \sqrt{8m\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right)}.$$

We further have (from London et al. (2016)) that

$$D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{j^\star}} \leq \frac{3}{2}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}\right).$$

Thus, with probability at least $1-\delta$, all posteriors satisfy

$$\begin{aligned}
&\mathop{\mathbb{E}}_{Z\sim\mathbb{D}}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\mathcal{L}(h,Z)] - \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\mathcal{L}(h,Z^{(i)})] \\
&\quad\le \frac{1}{\epsilon_{j^\star}}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{1}{\delta_{j^\star}}\right) + \frac{\epsilon_{j^\star}}{8m} \\
&\quad\le \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(1/\delta_{j^\star})}{\sqrt{2m(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta))}} + \frac{\sqrt{8m(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta))}}{8m} \\
&\quad\le \frac{3\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta)\right)}{2\sqrt{2m(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta))}} + \frac{\sqrt{8m(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta))}}{8m} \\
&\quad= 2\sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln(2/\delta)}{2m}},
\end{aligned}$$

which completes the proof. ■

B.1. Proof of Theorem 1

The proof proceeds in three stages: First, we construct prior and posterior distributions. Both are uniform, but the posterior is centered at a learned hypothesis, and its support shrinks as the training data grows. This construction essentially means that we should trust the learned model more as we get more training data. In the second step, we upper-bound the KL divergence term—which is trivial given the uniform constructions. Finally, we show that, under the posterior construction, the loss of a random hypothesis is "close to" the loss of the learned hypothesis; specifically, that the difference between their losses is a small additive term that vanishes as the training set grows.

Let $\mathbb{P}$ denote a uniform prior over $\{w \in \mathbb{R}^d : \|w\|_2 \le R\}$. Given a learned weight vector, w, we construct a posterior, $\mathbb{Q}$, as a uniform distribution over $\{w' \in \mathbb{R}^d : \|w'\|_2 \le R,\ \|w'-w\|_2 \le 2\gamma/(mB)\}$. It can easily be shown (London et al., 2016, Appendix C.3) that

$$D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) \le d\ln(mBR/\gamma). \tag{37}$$

The next part of the proof "derandomizes" the loss functions in Equation 30, replacing the randomized hypothesis, with weights $w' \sim \mathbb{Q}$, with the deterministic predictor based on w. To do so, we will bound the difference $|\mathcal{L}_\gamma(w,x) - \mathcal{L}_\gamma(w',x)|$. We first note that the ramp function is $(1/\gamma)$-Lipschitz; that is, for any γ, x, w and w',

$$\left|\mathcal{L}_\gamma(w,x) - \mathcal{L}_\gamma(w',x)\right| \le \frac{1}{\gamma}\left|D(w,x) - D(w',x)\right|. \tag{38}$$

To upper-bound $|D(w,x) - D(w',x)|$, let

$$\begin{aligned}
\mu_F &\in \mathop{\arg\max}_{\mu\in\mathcal{V}_F} \theta(x;w)^\top\mu \quad\text{and}\quad \mu'_F \in \mathop{\arg\max}_{\mu\in\mathcal{V}_F} \theta(x;w')^\top\mu, \\
\mu_I &\in \mathop{\arg\max}_{\mu\in\mathcal{V}_I} \theta(x;w)^\top\mu \quad\text{and}\quad \mu'_I \in \mathop{\arg\max}_{\mu\in\mathcal{V}_I} \theta(x;w')^\top\mu,
\end{aligned}$$

and observe that

$$\begin{aligned}\left|D(w,x) - D(w',x)\right| &= \left|\theta(x;w)^\top(\mu_F - \mu_I) - \theta(x;w')^\top(\mu'_F - \mu'_I)\right| \\ &\leq \left|\theta(x;w)^\top\mu_F - \theta(x;w')^\top\mu'_F\right| \qquad (39) \\ &\quad + \left|\theta(x;w')^\top\mu'_I - \theta(x;w)^\top\mu_I\right|. \qquad (40)\end{aligned}$$

We will upper-bound Equations 39 and 40 separately.

Starting with Equation 39, assume that $\theta(x;w)^\top\mu_F \geq \theta(x;w')^\top\mu'_F$. (If the inequality goes in the other direction, we simply swap the left and right terms, which is equivalent inside the absolute value.) We then have that

$$\begin{aligned}\left|\theta(x;w)^\top\mu_F - \theta(x;w')^\top\mu'_F\right| &= \theta(x;w)^\top\mu_F - \theta(x;w')^\top\mu'_F \\ &\leq \theta(x;w)^\top\mu_F - \theta(x;w')^\top\mu_F \\ &= (w-w')^\top\phi(x,\mu_F),\end{aligned}$$

due to the optimality of μ'_F for $\theta(x;w')$. Then, using Cauchy-Schwarz,

$$(w-w')^\top\phi(x,\mu_F) \leq \left\|w-w'\right\|_2 \left\|\phi(x,\mu_F)\right\|_2 \leq \left\|w-w'\right\|_2 B.$$

An identical analysis of Equation 40 yields

$$\begin{aligned}\left|\theta(x;w')^\top\mu'_I - \theta(x;w)^\top\mu_I\right| &= \theta(x;w')^\top\mu'_I - \theta(x;w)^\top\mu_I \\ &\leq \theta(x;w')^\top\mu'_I - \theta(x;w)^\top\mu'_I \\ &= (w'-w)^\top\phi(x,\mu'_I) \\ &\leq \left\|w'-w\right\|_2 \left\|\phi(x,\mu'_I)\right\|_2 \\ &\leq \left\|w'-w\right\|_2 B.\end{aligned}$$

By construction, every $w' \sim \mathbb{Q}$ has distance at most $2\gamma/(mB)$ from w. Therefore, substituting the above inequalities into Equations 39 and 40, we have that

$$\left|D(w,x) - D(w',x)\right| \leq \left\|w-w'\right\|_2 B + \left\|w'-w\right\|_2 B \leq \frac{2\gamma}{mB}\cdot 2B = \frac{4\gamma}{m};$$

and combining this bound with Equation 38 yields

$$\left|\mathcal{L}_\gamma(w,x) - \mathcal{L}_\gamma(w',x)\right| \leq \frac{1}{\gamma}\cdot\frac{4\gamma}{m} = \frac{4}{m}.$$

Thus, the loss of any random weight vector, w', is at most $4/m$ above or below that of the learned weights, w. We can therefore derandomize the randomized loss by bounding its distance to the deterministic loss:

$$\left|\mathcal{L}_\gamma(w,x) - \mathop{\mathbb{E}}_{w'\sim\mathbb{Q}}\left[\mathcal{L}_\gamma(w',x)\right]\right| \leq \mathop{\mathbb{E}}_{w'\sim\mathbb{Q}}\left[\left|\mathcal{L}_\gamma(w,x) - \mathcal{L}_\gamma(w',x)\right|\right] \leq \frac{4}{m}.$$

Recalling $\mathcal{L}_0(w,x) \le \mathcal{L}_\gamma(w,x)$, we then have that

$$\mathop{\mathbb{E}}_{x\sim\mathbb{D}}[\mathcal{L}_0(w,x)] \le \mathop{\mathbb{E}}_{x\sim\mathbb{D}}[\mathcal{L}_\gamma(w,x)] \le \mathop{\mathbb{E}}_{x\sim\mathbb{D}}\mathop{\mathbb{E}}_{w'\sim\mathbb{Q}}[\mathcal{L}_\gamma(w',x)] + \frac{4}{m} \tag{41}$$

$$\text{and} \quad \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{w'\sim\mathbb{Q}}[\mathcal{L}_\gamma(w',x^{(i)})] \le \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_\gamma(w,x^{(i)}) + \frac{4}{m}. \tag{42}$$

All that remains is to put the pieces together. Via Lemma 3, with probability at least $1-\delta$ over draws of the training data,

$$\mathop{\mathbb{E}}_{x\sim\mathbb{D}}\mathop{\mathbb{E}}_{w'\sim\mathbb{Q}}[\mathcal{L}_\gamma(w',x)] \le \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{w'\sim\mathbb{Q}}[\mathcal{L}_\gamma(w',x^{(i)})] + 2\sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2}{\delta}}{2m}}.$$

To complete the proof, we use Equation 37 to upper-bound the KL divergence, Equation 41 to lower-bound the randomized risk and Equation 42 to upper-bound the empirical randomized risk.

B.2. Proof of Theorem 2

As we did in the previous proof, we define $\mathbb{P}$ as a uniform prior over $\{w \in \mathbb{R}^d : \|w\|_2 \le R\}$. Given w, we construct a posterior, $\mathbb{Q}$, as a uniform distribution over $\{w' \in \mathbb{R}^d : \|w'\|_2 \le R,\ \|w'-w\|_2 \le 2/(mB)\}$ (we omit γ), which yields $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) \le d\ln(mBR)$.

In what follows we use the fact that when one of the arguments of $D_1(\mu,\mu')$ (Eq. (18))—say, μ—is integral, $D_1(\mu,\mu')$ can be expressed as $\ell(\mu)^\top\mu'$, where

$$\ell(\mu) \triangleq \frac{1}{n}\begin{bmatrix}\mathbf{1}-\mu_{\mathrm{u}}\\ \mathbf{0}\end{bmatrix}. \tag{43}$$

To derandomize the loss of $w' \sim \mathbb{Q}$, we need to bound the difference $|\mathcal{L}_{\mathrm{r}}(w,x,\mu_I) - \mathcal{L}_{\mathrm{r}}(w',x,\mu_I)|$. Note that w' is being evaluated with respect to the MAP assignment under w, i.e., $\mu_I \in \arg\max_{\mu\in\mathcal{M}}\theta(x;w)^\top\mu$. To simplify notation, we will use the following shorthand:

$$\theta \triangleq \theta(x;w) \quad \text{and} \quad \theta' \triangleq \theta(x;w')\,;$$

$$\mu_L \in \mathop{\arg\max}_{\mu\in\mathcal{M}_L}\ \theta^\top\mu \quad \text{and} \quad \mu'_L \in \mathop{\arg\max}_{\mu\in\mathcal{M}_L}\ (\theta')^\top\mu\,;$$

$$\tilde{\mu}_L \in \mathop{\arg\max}_{\mu\in\mathcal{M}_L}\ (\ell(\mu_I)+\theta)^\top\mu \quad \text{and} \quad \tilde{\mu}'_L \in \mathop{\arg\max}_{\mu\in\mathcal{M}_L}\ (\ell(\mu_I)+\theta')^\top\mu.$$

We then have that the difference of ramp losses decomposes as

$$\begin{aligned}
&\big|\mathcal{L}_{\mathrm{r}}(w,x,\mu_I) - \mathcal{L}_{\mathrm{r}}(w',x,\mu_I)\big| \\
&\quad= \left|\Big((\ell(\mu_I)+\theta)^\top\tilde{\mu}_L - \theta^\top\mu_L\Big) \ - \ \Big((\ell(\mu_I)+\theta')^\top\tilde{\mu}'_L - (\theta')^\top\mu'_L\Big)\right| \\
&\quad\le \left|(\ell(\mu_I)+\theta)^\top\tilde{\mu}_L - (\ell(\mu_I)+\theta')^\top\tilde{\mu}'_L\right| &(44)\\
&\qquad+ \left|(\theta')^\top\mu'_L - \theta^\top\mu_L\right|. &(45)
\end{aligned}$$

Just as before, we will upper-bound Equations 44 and 45 separately.

We start with Equation 45, which is slightly simpler. Without loss of generality, assume that $(\theta')^\top \mu'_L \geq \theta^\top \mu_L$. Using the optimality of μ_L with respect to θ, we then have that

$$\begin{aligned}\left|(\theta')^\top \mu'_L - \theta^\top \mu_L\right| &= (\theta')^\top \mu'_L - \theta^\top \mu_L \\ &\leq (\theta')^\top \mu'_L - \theta^\top \mu'_L \\ &= (w' - w)^\top \phi(x, \mu'_L) \\ &\leq \left\|w' - w\right\|_2 \left\|\phi(x, \mu'_L)\right\|_2 \\ &\leq \left\|w' - w\right\|_2 B.\end{aligned}$$

By construction, every $w' \sim \mathbb{Q}$ has distance at most $2/(mB)$ from w, so

$$\left|(\theta')^\top \mu'_L - \theta^\top \mu_L\right| \leq \left\|w' - w\right\|_2 B \leq \frac{2}{mB} \cdot B = \frac{2}{m}. \tag{46}$$

Applying the same technique to Equation 44, we assume, without loss of generality, that $(\ell(\mu_I) + \theta)^\top \tilde{\mu}_L \geq (\ell(\mu_I) + \theta')^\top \tilde{\mu}'_L$. Then,

$$\begin{aligned}\left|(\ell(\mu_I) + \theta)^\top \tilde{\mu}_L - (\ell(\mu_I) + \theta')^\top \tilde{\mu}'_L\right| &= (\ell(\mu_I) + \theta)^\top \tilde{\mu}_L - (\ell(\mu_I) + \theta')^\top \tilde{\mu}'_L \\ &\leq (\ell(\mu_I) + \theta)^\top \tilde{\mu}_L - (\ell(\mu_I) + \theta')^\top \tilde{\mu}_L \\ &= (\theta - \theta')^\top \tilde{\mu}_L \\ &= (w - w')^\top \phi(x, \tilde{\mu}_L) \\ &\leq \left\|w - w'\right\|_2 B \leq \frac{2}{m}.\end{aligned} \tag{47}$$

Substituting Equation 47 for 44, and Equation 46 for 45, we have that

$$\left|\mathcal{L}_{\mathrm{r}}(w, x, \mu_I) - \mathcal{L}_{\mathrm{r}}(w', x, \mu_I)\right| \leq \frac{4}{m} \ ;$$

hence,

$$\left|\mathcal{L}_{\mathrm{r}}(w, x, \mu_I) - \mathop{\mathbb{E}}_{w' \sim \mathbb{Q}}[\mathcal{L}_{\mathrm{r}}(w', x, \mu_I)]\right| \leq \mathop{\mathbb{E}}_{w' \sim \mathbb{Q}}\left[\left|\mathcal{L}_{\mathrm{r}}(w, x, \mu_I) - \mathcal{L}_{\mathrm{r}}(w', x, \mu_I)\right|\right] \leq \frac{4}{m}.$$

Then, using the inequalities in Equation 22, we have that

$$\mathop{\mathbb{E}}_{x \sim \mathbb{D}}[\mathcal{L}_1(w, x, \mu_I)] \leq \mathop{\mathbb{E}}_{x \sim \mathbb{D}}[\mathcal{L}_{\mathrm{r}}(w, x, \mu_I)] \leq \mathop{\mathbb{E}}_{x \sim \mathbb{D}} \mathop{\mathbb{E}}_{w' \sim \mathbb{Q}}[\mathcal{L}_{\mathrm{r}}(w', x, \mu_I)] + \frac{4}{m}, \tag{48}$$

and

$$\frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{w' \sim \mathbb{Q}}[\mathcal{L}_{\mathrm{r}}(w', x^{(i)}, \mu_I^{(i)})] \leq \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_{\mathrm{r}}(w, x^{(i)}, \mu_I^{(i)}) + \frac{4}{m}. \tag{49}$$

Finally, combining Lemma 3 with Equations 48 and 49 to lower- and upper-bound the randomized losses, we have that, with probability at least $1 - \delta$ over draws of the training

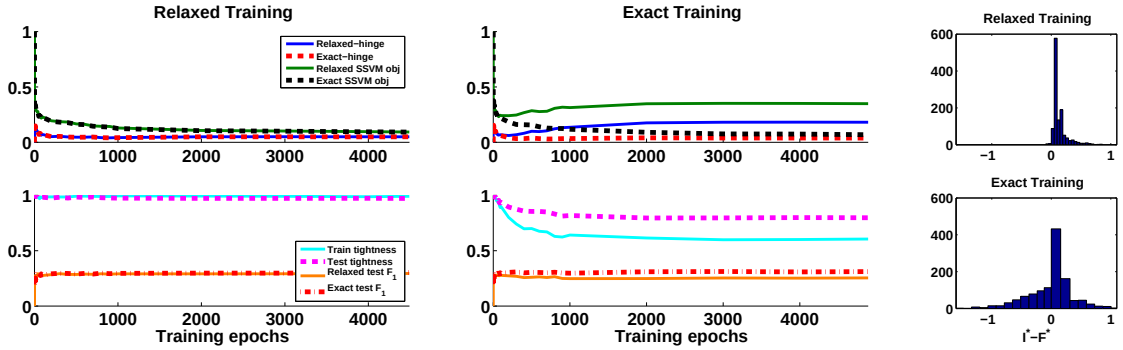

Figure 7: Training with a noisy version of the 'Scene' dataset. Various quantities of interest are shown as a function of training iterations. (Left) Training with LP relaxation. (Middle) Training with ILP. (Right) Integrality margin (bin widths are scaled differently).

set,

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathbb{D}}[\mathcal{L}_1(w,x,\mu_I)] &\leq \mathop{\mathbb{E}}_{x\sim\mathbb{D}}\mathop{\mathbb{E}}_{w'\sim\mathbb{Q}}[\mathcal{L}_{\mathrm{r}}(w',x,\mu_I)] + \frac{4}{m} \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{w'\sim\mathbb{Q}}[\mathcal{L}_{\mathrm{r}}(w',x^{(i)},\mu_I^{(i)})] + \frac{4}{m} + 2\sqrt{\frac{d\ln(mBR)+\ln\frac{2}{\delta}}{2m}} \\
&\leq \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_{\mathrm{r}}(w,x^{(i)},\mu_I^{(i)}) + \frac{8}{m} + 2\sqrt{\frac{d\ln(mBR)+\ln\frac{2}{\delta}}{2m}}.
\end{aligned}$$

Noting that the hinge loss uniformly upper-bounds the ramp loss completes the proof.

Appendix C. Additional Experimental Results

In this section we present additional experimental results for the 'Scene' dataset. Specifically, we inject random Gaussian noise to the input features in order to reduce the signal in the singleton scores and increase the role of the pairwise interactions. This makes the problem harder since the prediction needs to account for global information.

In Fig. 7 we observe that with exact training the exact loss is minimized, causing the exact-hinge to decrease, since it is upper bounded by the loss (middle, top). On the other hand, the relaxed-hinge (and relaxed loss) *increase* during training, which results in a large integrality gap and fewer tight instances. In contrast, with relaxed training the relaxed loss is minimized, which causes the relaxed-hinge to decrease. Since the exact-hinge is upper bounded by the relaxed-hinge it also decreases, but both hinge terms decrease similarly and remain very close to each other. This results in a small integrality gap and tightness of almost all instances.

Finally, in contrast to other settings, in Fig. 7 we observe that with exact training the test tightness is noticeably higher (about 20%) than the train tightness (Fig. 7, middle, bottom).

This does not contradict our bound from Theorem 1, since in fact the test fractionality is even *lower* than the bound suggests. On the other hand, this result does indicate that train and test tightness may sometimes behave differently, which means that we might need to increase the size of the trainset in order to get a tighter bound.

References

G. H. Bakir, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan. *Predicting Structured Data*. The MIT Press, 2007.

F. Barahona. On cuts and matchings in planar graphs. *Mathematical Programming*, 60: 53–68, 1993.

E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *CVPR*, 2004.

C. Chekuri, S. Khanna, J. Naor, and L. Zosin. A linear programming formulation and approximation algorithms for the metric labeling problem. *SIAM J. on Discrete Mathematics*, 18(3):608–625, 2004.

L.-C. Chen, A. Schwing, A. Yuille, and R. Urtasun. Learning deep structured models. In *International Conference on Machine Learning*, 2015.

M. Collins. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *EMNLP*, 2002.

Hal Daumé III, John Langford, and Daniel Marcu. Search-based structured prediction. *Machine Learning*, 75(3):297–325, 2009.

J. Domke. Learning graphical model parameters with approximate marginal inference. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(10), 2013.

M. Donsker and S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

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, 2008.

A. Globerson, T. Roughgarden, D. Sontag, and C. Yildirim. How hard is inference for structured prediction? In *ICML*, 2015.

J. Kleinberg and E. Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields. *Journal of the ACM*, 49(5):616–639, 2002.

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

A. Koster, S.P.M. van Hoesel, and A.W.J. Kolen. The partial constraint satisfaction problem: Facets and lifting theorems. *Operations Research Letters*, 23:89–97, 1998.

A. Kulesza and F. Pereira. Structured learning with approximate inference. In *Advances in Neural Information Processing Systems 20*, pages 785–792, 2007.

S. Lacoste-Julien, B. Taskar, D. Klein, and M. I. Jordan. Word alignment via quadratic assignment. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 112–119, 2006.

S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, pages 53–61, 2013.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Optimality of approximate inference algorithms on stable instances. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS)*. JMLR: W&CP, 2018.

D. Khuê Lê-Huu and Nikos Paragios. Continuous Relaxation of MAP Inference: A Nonconvex Perspective. In *CVPR 2018 - IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–19, 2018.

B. London, B. Huang, and L. Getoor. Stability and generalization in structured prediction. *Journal of Machine Learning Research*, 17, 2016.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Bilu–Linial stable instances of max cut and minimum multiway cut. *Proc. 22nd Symposium on Discrete Algorithms (SODA)*, 2014.

A. Martins, N. Smith, and E. P. Xing. Concise integer linear programming formulations for dependency parsing. In *ACL*, 2009a.

A. Martins, N. Smith, and E. P. Xing. Polyhedral outer approximations with application to natural language parsing. In *Proceedings of the 26th International Conference on Machine Learning*, 2009b.

D. McAllester. Generalization bounds and consistency for structured labeling. In G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, editors, *Predicting Structured Data*. MIT Press, 2007.

O. Meshi, E. Eban, G. Elidan, and A. Globerson. Learning max-margin tree predictors. In *UAI*, 2013.

Sebastian Nowozin, Peter V. Gehler, Jeremy Jancsary, and Christoph Lampert. *Advanced Structured Prediction*. MIT Press, 2014.

Manfred Padberg. The boolean quadric polytope: Some characteristics, facets and relatives. *Mathematical Programming*, 45(1):139–172, 1989.

P. Raghavan and C. Tompson. Randomized rounding: A technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7(4):365–374, 1987.

P. Ravikumar, A. Agarwal, and M. J. Wainwright. Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *JMLR*, 11:1043–1080, 2010.

D. Roth. On the hardness of approximate reasoning. *Artificial Intelligence*, 82, 1996.

D. Roth and W. Yih. A linear programming formulation for global inference in natural language tasks. In *CoNLL, The 8th Conference on Natural Language Learning*, 2004.

D. Roth and W. Yih. Integer linear programming inference for conditional random fields. In *ICML*, pages 736–743. ACM, 2005.

M. Rowland, A. Pacchiano, and A. Weller. Conditions beyond treewidth for tightness of higher-order LP relaxations. In *In Artificial Intelligence and Statistics (AISTATS)*, 2017.

Alexander M. Rush, David Sontag, Michael Collins, and Tommi Jaakkola. On dual decomposition and linear programming relaxations for natural language processing. In *EMNLP*, 2010.

M. I. Schlesinger. Syntactic analysis of two-dimensional visual signals in noisy conditions. *Kibernetika*, 4:113—130, 1976.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

H. D. Sherali and W. P. Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J. on Disc. Math.*, 3(3):411–430, 1990.

Y. Shimony. Finding the MAPs for belief networks is NP-hard. *Aritifical Intelligence*, 68(2): 399–410, 1994.

D. Sontag and T. Jaakkola. New outer bounds on the marginal polytope. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1393–1400. MIT Press, Cambridge, MA, 2008.

D. Sontag, T. Meltzer, A. Globerson, T. Jaakkola, and Y. Weiss. Tightening LP relaxations for MAP using message passing. In *UAI*, pages 503–510, 2008.

D. Sontag, O. Meshi, T. Jaakkola, and A. Globerson. More data means less inference: A pseudo-max approach to structured learning. In *NIPS*, 2010.

B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems*. MIT Press, 2003.

B. Taskar, V. Chatalbashev, and D. Koller. Learning associative Markov networks. In *Proc. ICML*. ACM Press, 2004.

J. Thapper and S. Živný. The power of linear programming for valued CSPs. In *FOCS*, 2012.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*, pages 104–112, 2004.

M. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.

M. Wainwright, T. Jaakkola, and A. Willsky. MAP estimation via agreement on trees: message-passing and linear programming. *IEEE Transactions on Information Theory*, 51 (11):3697–3717, 2005.

D. Weiss and B. Taskar. Structured Prediction Cascades. In *AISTATS*, 2010.

A. Weller. Characterizing tightness of LP relaxations by forbidding signed minors. In *Uncertainty in Artificial Intelligence (UAI)*, 2016.

Adrian Weller. Bethe and related pairwise entropy approximations. In *Uncertainty in Artificial Intelligence (UAI)*, 2015.

Adrian Weller, Mark Rowland, and David Sontag. Tightness of LP relaxations for almost balanced models. In *AISTATS*, 2016.

Yuan Zhang, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Greed is good if randomized: New inference for dependency parsing. In *EMNLP*, 2014.