On the Tightness of LP Relaxations for Structured Prediction

Ofer Meshi **Mehrdad Mahdavi**
Toyota Technological Institute at Chicago

David Sontag
Courant Institute of Mathematical Sciences
New York University

Abstract

Structured prediction applications often involve complex inference problems that require the use of approximate methods. Approximations based on linear programming (LP) relaxations have proved particularly successful in this setting, with both theoretical and empirical support. Despite the general intractability of inference, it has been observed that in many real-world applications the LP relaxation is often tight. In this work we propose a theoretical explanation to this striking observation. In particular, we show that learning with LP relaxed inference encourages tightness of training instances. We complement this result with a generalization bound showing that tightness generalizes from train to test data.

1 Introduction

Many applications of machine learning can be formulated as prediction problems over structured output spaces (Bakir et al., 2007). 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. 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 quite 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 (an *integral* solution is found). This is particularly surprising since some of these LPs are quite complicated, consisting of thousands of variables and constraints, and their score function is not constrained in any way. So why are these real-world instances not as bad as their theoretical worst case?

Unfortunately, our understanding of this interesting phenomenon has been lacking. This work aims to address this question and provide a theoretical explanation for the tightness of LP relaxations in the context of structured prediction. In particular, we show that the 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 does not have a similar effect, which is consistent with previous empirical findings. To explain tightness of *test* instances, we complement this result with a generalization bound for integrality. Our bound implies that if many training instances are integral, then test instances are also likely to be integral.

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 $y = (y_1, \ldots, y_n)$. A popular model class for this task is based on linear classifiers. In this setting prediction is performed via a linear discriminant rule: $y(x;w) = \operatorname{argmax}_{y'} w^\top \phi(x,y')$, 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 c. 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: $\max_y \sum_c \theta_c(y_c;x,w)$ (we will sometimes omit the dependence on x and w in the sequel).

Due to its combinatorial nature, the prediction problem remains NP-hard despite the decomposition assumption. Fortunately, efficient approximations have been proposed. Here we will be particularly interested in approximations based on LP relaxations. We begin by formulating prediction as the following ILP:[1]

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

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

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). 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. Depending on the scores θ, sometimes the optimal solution to the relaxed LP may actually be integral (i.e., satisfy $\mu \in \{0,1\}^q$). In fact, for some score functions, such as super-modular scores, this is guaranteed to happen (Taskar et al., 2004). For a more general characterization see Thapper and Živný (2012). However, these sufficient conditions are by no means necessary, and indeed, many score functions that are useful in practice do not satisfy them but still produce highly integral solutions (Sontag et al., 2008; Finley and Joachims, 2008; Martins et al., 2009; Koo et al., 2010). In the next section we propose a simple and intuitive explanation for this empirical observation.

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 $\Delta(y(x^{(m)};w), y^{(m)})$. 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:

$$\min_w \sum_m \max_y \left[w^\top \Big(\phi(x^{(m)}, y) - \phi(x^{(m)}, y^{(m)}) \Big) + \Delta(y, y^{(m)}) \right] . \tag{1}$$

This 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. Fortunately, as in prediction, LP relaxation can be applied to the structured loss (Taskar et al., 2003; Kulesza and Pereira, 2007), which yields the relaxed training objective:

$$\min_w \sum_m \max_{\mu \in \mathcal{M}_L} \left[\theta_m^\top (\mu - \mu_m) + \ell_m^\top \mu \right] , \tag{2}$$

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, similarly $\ell_m \in \mathbb{R}^q$ is a vector with entries $\Delta_c(y_c, y_c^{(m)})$, and μ_m is the integral vector corresponding to $y^{(m)}$.

[1] For convenience we introduce singleton factors θ_i, which can be set to 0 if needed.

3 An Argument for Integrality

In this section we present our main result, proposing an explanation for the observed tightness of LP relaxations. To this end, we make two complementary arguments: in Section 3.1 we argue that optimizing the relaxed training objective of Eq. (2) also has the effect of encouraging integrality of training instances; in Section 3.2 we show that integrality generalizes from train to test data.

3.1 Integrality at Training

We first show that the *relaxed* training objective in Eq. (2), although designed to achieve high accuracy, also induces integrality. 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:

$$\mu_L = \operatorname*{argmax}_{\mu \in \mathcal{M}_L} \theta^\top \mu \qquad\qquad \mu_I = \operatorname*{argmax}_{\substack{\mu \in \mathcal{M}_L \\ \mu \in \{0,1\}^q}} \theta^\top \mu$$

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

$$\underbrace{\theta^\top(\mu_L - \mu_T)}_{\text{relaxed-hinge}} = \underbrace{\theta^\top(\mu_L - \mu_I)}_{\text{integrality gap}} + \underbrace{\theta^\top(\mu_I - \mu_T)}_{\text{exact-hinge}} \tag{3}$$

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 terms are non-negative). This simple decomposition has several interesting implications.

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

$$\begin{aligned}
\theta^\top(\mu_L - \mu_I) &= \theta^\top(\mu_L - \mu_T) - \theta^\top(\mu_I - \mu_T) && (4)\\
&\le \theta^\top(\mu_L - \mu_T) && (5)\\
&\le \theta^\top(\mu_L - \mu_T) + \ell^\top \mu_L && (6)\\
&\le \max_{\mu \in \mathcal{M}_L} \left(\theta^\top(\mu - \mu_T) + \ell^\top \mu\right) && (7)
\end{aligned}$$

Where Eq. (7) is precisely the relaxed training objective from Eq. (2). Therefore, optimizing the approximate training objective of Eq. (2) *minimizes an upper bound on the integrality gap*. Hence, driving down the approximate objective also reduces the integrality gap of training instances. One case where the integrality gap becomes zero is when the data is algorithmically separable (i.e., the approximate loss is 0). In this case the relaxed-hinge term vanishes (the exact-hinge must also vanish), and integrality is assured. However, the bound above might sometimes be loose. At the same time, Eq. (4) 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, we can still obtain a small gap 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?

To get a better understanding, a key insight is that the relaxed and exact hinge terms are closely related to the relaxed and exact training objectives (the latter additionally depend on the task loss Δ). Intuitively, minimizing the training objective also minimizes the corresponding hinge term. Using this intuition, we realize that with relaxed training, the relaxed-hinge is reduced, which consequently reduces the exact-hinge (since it is an upper-bound). However, the training objective does not encourage reduction in the exact-hinge directly, so it is reasonable to expect that it would not be much smaller than the relaxed-hinge. Therefore, relaxed training is likely to induce a small integrality gap. We would like to point out that this kind of behavior is not guaranteed, and our explanation merely serves to provide an intuition to what happens in practice. Indeed, it is possible to construct a learning scenario where relaxed training obtains zero exact-hinge and non-zero relaxed-hinge, so the relaxation is not tight in that case.

In contrast to relaxed training, the same intuition suggests that *exact training may actually increase the integrality gap*. This might occur since it minimizes the exact-hinge (upper bounded by the exact training objective) without also reducing the relaxed-hinge directly. This intuition is consistent with previous empirical evidence. Specifically, Martins et al. (2009, Table 2) showed that training with relaxed objective achieved 92.88% integral solutions, while exact training achieved only 83.47% integral solutions on a dependency parsing problem. An even stronger effect was observed by Finley and Joachims (2008, Table 3) for multi-label classification, where relaxed training resulted in 99.57% integral instances, with exact training attaining only 17.7% ('Yeast' dataset).

Finally, note that our derivation above (Eq. (4)) holds for *any integral* μ, and not just the ground-truth μ_T. In other words, the only property of μ_T we are using here is its integrality. Indeed, we verify empirically (not shown) that training a model using *random labels* still attains the same level of integrality as training with the ground-truth labels. This analysis suggests that *integrality is not related to accuracy* of the predictor.

3.2 Generalization of Integrality

Our argument in Section 3.1 concerns only the integrality of train instances. However, the empirical evidence discussed above pertains to test data. To bridge this gap, in this section we show that train integrality implies test integrality. We do so by proving a generalization bound for integrality based on Rademacher complexity.

We first define a loss function which measures the lack of integrality (or, fractionality). To this end, we consider the discrete set of *vertices* of the local polytope $\mathcal{M}_L$ (excluding its convex hull), denoting by $\mathcal{M}^I$ and $\mathcal{M}^F$ the sets of fully-integral and non-integral (i.e., fractional) vertices, respectively (so $\mathcal{M}^I \cap \mathcal{M}^F = \emptyset$, and $\mathcal{M}^I \cup \mathcal{M}^F$ consists of all vertices of $\mathcal{M}_L$). Next, let $\theta_x \in \mathbb{R}^q$ be the mapping from weights w and inputs x to scores (as used in Eq. (2)), and let $I^*(\theta) = \max_{\mu \in \mathcal{M}^I} \theta^\top \mu$ and $F^*(\theta) = \max_{\mu \in \mathcal{M}^F} \theta^\top \mu$ be the best integral and fractional scores attainable, respectively. The fractionality of θ can be measured by the quantity[2] $D = F^* - I^*$. If this quantity is large then the LP has a fractional solution with a much better score than any integral solution. We can now define the loss:

$$\mathcal{L}(\theta) = \begin{cases} 1 & D > 0 \\ 0 & \text{otherwise} \end{cases} . \tag{8}$$

That is, the loss equals 1 if and only if the optimal fractional solution has a (strictly) higher score than the optimal integral solution.[3] In addition, we define a ramp loss parameterized by $\gamma > 0$ which upper bounds the fractionality loss:

$$\varphi_\gamma(\theta) = \begin{cases} 0 & D \le -\gamma \\ 1 + D/\gamma & -\gamma < D \le 0 \\ 1 & D > 0 \end{cases} , \tag{9}$$

Notice that for this loss to be zero, the best integral solution has to be better than the best fractional solution by at least γ, which is a stronger requirement than mere tightness. We also point out that $\varphi_\gamma(\theta)$ is generally hard to compute, as is $\mathcal{L}(\theta)$ (due to the discrete optimization involved in computing I^* and F^*). However, here we are only interested in proving that integrality is a generalizing property, so we will not worry about computational efficiency for now. We are now ready to state the main theorem of this section (the proof is similar in spirit to the one in Weiss and Taskar (2010), and relies on a general result from Bartlett and Mendelson (2002)).

Theorem 3.1. *Let inputs be independently selected according to a probability measure $P(X)$, and let Θ be the class of all scoring functions θ_X with $\|w\|_2 \le B$. Let $\|\phi(x, y_c)\|_2 \le \hat{R}$ for all x, c, y_c, and q is the total number of factor assignments (dimension of μ). Then for any number of samples M and any $0 < \delta < 1$, with probability at least $1 - \delta$, every $\theta_X \in \Theta$ satisfies:*

$$\mathbb{E}_P[\mathcal{L}(\theta_X)] \le \hat{\mathbb{E}}_M[\varphi_\gamma(\theta_X)] + O\left(\frac{q^{1.5} B \hat{R}}{\gamma \sqrt{M}}\right) + \sqrt{\frac{8 \ln(2/\delta)}{M}} \tag{10}$$

where $\hat{\mathbb{E}}_M$ is the empirical expectation. This theorem suggests 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 as the number of samples grows.

[2] In order to simplify notation we omit the dependence of I^*, F^*, and D on θ.

[3] Notice that the loss will be 0 whenever the non-integral and integral optima are equal, but this is fine for our purpose, since we consider the relaxation to be tight in this case.

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.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2002.

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.

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

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*, 2009.

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.

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.

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