Structured Prediction Theory Based on Factor Graph Complexity

Authored by:

Corinna Cortes
Mehryar Mohri
Scott Yang
Vitaly Kuznetsov

Abstract

We present a general theoretical analysis of structured prediction with a series of new results. We give new data-dependent margin guarantees for structured prediction for a very wide family of loss functions and a general family of hypotheses, with an arbitrary factor graph decomposition. These are the tightest margin bounds known for both standard multi-class and general structured prediction problems. Our guarantees are expressed in terms of a data-dependent complexity measure, emph{factor graph complexity}, which we show can be estimated from data and bounded in terms of familiar quantities for several commonly used hypothesis sets, and a sparsity measure for features and graphs. Our proof techniques include generalizations of Talagrand's contraction lemma that can be of independent interest. We further extend our theory by leveraging the principle of Voted Risk Minimization (VRM) and show that learning is possible even with complex factor graphs. We present new learning bounds for this advanced setting, which we use to devise two new algorithms, emph{Voted Conditional Random Field} (VCRF) and emph{Voted Structured Boosting} (StructBoost). These algorithms can make use of complex features and factor graphs and yet benefit from favorable learning guarantees. We also report the results of experiments with VCRF on several datasets to validate our theory.

1 Paper Body

Structured prediction covers a broad family of important learning problems. These include key tasks in natural language processing such as part-of-speech tagging, parsing, machine translation, and named-entity recognition, important areas in computer vision such as image segmentation and object recognition, and also crucial areas in speech processing such as pronunciation modeling and

speech recognition. In all these problems, the output space admits some structure. This may be a sequence of tags as in part-of-speech tagging, a parse tree as in context-free parsing, an acyclic graph as in dependency parsing, or labels of image segments as in object detection. Another property common to these tasks is that, in each case, the natural loss function admits a decomposition along the output substructures. As an example, the loss function may be the Hamming loss as in part-of-speech tagging, or it may be the edit-distance, which is widely used in natural language and speech processing. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

The output structure and corresponding loss function make these problems significantly different from the (unstructured) binary classification problems extensively studied in learning theory. In recent years, a number of different algorithms have been designed for structured prediction, including Conditional Random Field (CRF) [Lafferty et al., 2001], StructSVM [Tsochantaridis et al., 2005], Maximum-Margin Markov Network (M3N) [Taskar et al., 2003], a kernel-regression algorithm [Cortes et al., 2007], and search-based approaches such as [Daum? III et al., 2009, Doppa et al., 2014, Lam et al., 2015, Chang et al., 2015, Ross et al., 2011]. More recently, deep learning techniques have also been developed for tasks including part-of-speech tagging [Jurafsky and Martin, 2009, Vinyals et al., 2015a], named-entity recognition [Nadeau and Sekine, 2007], machine translation [Zhang et al., 2008], image segmentation [Lucchi et al., 2013], and image annotation [Vinyals et al., 2015b]. However, in contrast to the plethora of algorithms, there have been relatively few studies devoted to the theoretical understanding of structured prediction [Bakir et al., 2007]. Existing learning guarantees hold primarily for simple losses such as the Hamming loss [Taskar et al., 2003, Cortes et al., 2014, Collins, 2001] and do not cover other natural losses such as the edit-distance. They also typically only apply to specific factor graph models. The main exception is the work of McAllester [2007], which provides PAC-Bayesian guarantees for arbitrary losses, though only in the special case of randomized algorithms using linear (count-based) hypotheses. This paper presents a general theoretical analysis of structured prediction with a series of new results. We give new data-dependent margin guarantees for structured prediction for a broad family of loss functions and a general family of hypotheses, with an arbitrary factor graph decomposition. These are the tightest margin bounds known for both standard multi-class and general structured prediction problems. For special cases studied in the past, our learning bounds match or improve upon the previously best bounds (see Section 3.3). In particular, our bounds improve upon those of Taskar et al. [2003]. Our guarantees are expressed in terms of a data-dependent complexity measure, factor graph complexity, which we show can be estimated from data and bounded in terms of familiar quantities for several commonly used hypothesis sets along with a sparsity measure for features and graphs. We further extend our theory by leveraging the principle of Voted Risk Minimization (VRM) and show that learning is possible even with complex factor graphs. We present new learning bounds for this advanced setting, which we use to design two new algorithms, Voted Conditional Random Field (VCRF) and Voted Structured Boosting (StructBoost).

These algorithms can make use of complex features and factor graphs and yet benefit from favorable learning guarantees. As a proof of concept validating our theory, we also report the results of experiments with VCRF on several datasets. The paper is organized as follows. In Section 2 we introduce the notation and definitions relevant to our discussion of structured prediction. In Section 3, we derive a series of new learning guarantees for structured prediction, which are then used to prove the VRM principle in Section 4. Section 5 develops the algorithmic framework which is directly based on our theory. In Section 6, we provide some preliminary experimental results that serve as a proof of concept for our theory.

2

Preliminaries

Let X denote the input space and Y the output space. In structured prediction, the output space may be a set of sequences, images, graphs, parse trees, lists, or some other (typically discrete) objects admitting some possibly overlapping structure. Thus, we assume that the output structure can be decomposed into l substructures. For example, this may be positions along a sequence, so that the output space Y is decomposable along these substructures: Y = Y1 ? ? ? ? ? Yl . Here, Yk is the set of possible labels (or classes) that can be assigned to substructure k. Loss functions. We denote by L : Y ? Y ! R+ a loss function measuring the dissimilarity of two elements of the output space Y. We will assume that the loss function L is definite, that is L(y, y 0) = 0 iff y = y 0 . This assumption holds for all loss functions commonly used in structured prediction. A key aspect of structured prediction is that the loss function can be decomposed along the Pl substructures Yk . As an example, L may be the Hamming loss defined by L(y, y 0) = 1l k=1 1yk 6=yk0 for all y = (y1 , . . . , yl) and y 0 = (y10 , . . . , yl0), with yk , yk0 2 Yk . In the common case where Y is a set of sequences defined over a finite alphabet, L may be the edit-distance, which is widely used in natural language and speech processing applications, with possibly different costs associated to insertions, deletions and substitutions. L may also be a loss based on the negative inner product of the vectors of n-gram counts of two sequences, or its negative logarithm. Such losses have been 2

1

f1

f2

2

3

2

f2

1

f1

(a)

3

(b)

Figure 1: Example of factor graphs. (a) Pairwise Markov network decomposition: h(x, y) = hf1 (x, y1 , y2) + hf2 (x, y2 , y3) (b) Other decomposition h(x, y) = hf1 (x, y1 , y3) + hf2 (x, y1 , y2 , y3). used to approximate the BLEU score loss in machine translation. There are other losses defined in computational biology based on various string-similarity measures. Our theoretical analysis is general and applies to arbitrary bounded and definite loss functions. Scoring functions and factor graphs. We will adopt the common approach in structured prediction where predictions are based on a scoring function mapping X ? Y to R. Let H be a family of scoring functions. For any h 2 H, we denote by h the predictor defined by h: for any x 2 X , h(x) = argmaxy2Y h(x, y). Furthermore, we will assume, as is standard in structured prediction, that each function h 2 H can be decomposed as a sum. We will consider the most general case for such decompositions, which can be made explicit using the notion of factor graphs.1 A factor graph G is a tuple G = (V, F, E), where V is a set of variable nodes, F a set of factor nodes, and E a set of undirected edges between a variable node and a factor node. In our context, V can be identified with the set of substructure indices, that is V = {1, . . . , l}. For any factor node f , denote by N(f) ? V the set of variable nodes connected to f via an edge and Q define Yf as the substructure set cross-product Yf = k2N(f) Yk . Then, h admits the following decomposition as a sum of functions hf , each taking as argument an element of the input space x 2 X and an element of Yf , yf 2 Yf : X h(x, y) = hf (x, yf). (1) f 2F

Figure 1 illustrates this definition with two different decompositions. More generally, we will consider the setting in which a factor graph may depend on a particular example (xi , yi): G(xi , yi) = Gi = ([li], Fi , Ei). A special case of this setting is for example when the size li (or length) of each example is allowed to vary and where the number of possible labels —Y— is potentially infinite.

We present other examples of such hypothesis sets and their decomposition in Section 3, where we discuss our learning guarantees. Note that such hypothesis sets H with an additive decomposition are those commonly used in most structured prediction algorithms [Tsochantaridis et al., 2005, Taskar et al., 2003, Lafferty et al., 2001]. This is largely motivated by the computational requirement for efficient training and inference. Our results, while very general, further provide a statistical learning motivation for such decompositions. Learning scenario. We consider the familiar supervised learning scenario where the training and test points are drawn i.i.d. according to some distribution D over X ? Y. We will further adopt the standard definitions of margin, generalization error and empirical error. The margin ?h (x, y) of a hypothesis h for a labeled example (x, y) 2 X ? Y is defined by ?h (x, y) = h(x, y)

max h(x, y 0). 0

(2)

y 6=y

Let S = ((x1 , y1), . . . , (xm , ym)) be a training sample of size m drawn from Dm . We denote by bS (h) the empirical error of h over S: R(h) the generalization error and by R R(h) =

E
(x,y)?D
[L(h(x), y)]
and
bS (h) = R
E
(x,y)?S
[L(h(x), y)],
(3)

1 Factor graphs are typically used to indicate the factorization of a probabilistic model. We are not assuming probabilistic models, but they would be also captured by our general framework: h would then be - log of a probability.

3

where h(x) = argmaxy h(x, y) and where the notation (x, y) ? S indicates that (x, y) is drawn according to the empirical distribution defined by S. The learning problem consists of using the sample S to select a hypothesis h 2 H with small expected loss R(h).

Observe that the definiteness of the loss function implies, for all x 2 X , the following equality: L(h(x), y) = L(h(x), y) 1?h (x,y)?0 . (4) We will later use this identity in the derivation of surrogate loss functions.

3

General learning bounds for structured prediction

In this section, we present new learning guarantees for structured prediction. Our analysis is general and applies to the broad family of definite and bounded loss functions described in the previous section. It is also general in the sense that it applies to general hypothesis sets and not just sub-families of linear functions. For linear hypotheses, we will give a more refined analysis that holds for arbitrary norm-p regularized hypothesis sets. The theoretical analysis of structured prediction is more complex than for classification since, by definition, it depends on the properties of the loss function and the factor graph. These attributes capture the combinatorial properties of the problem which must be exploited since the total number of labels is often exponential in the size of that graph. To tackle this problem, we first introduce a new complexity tool. 3.1

Complexity measure

A key ingredient of our analysis is a new data-dependent notion of complexity that extends the classical Rademacher complexity. We define the empirical factor graph Rademacher complexity b G (H) of a hypothesis set H for a sample S = (x1 , . . . , xm) and factor graph G as follows: R S " # m X X p X 1 G b (H) = R E sup —Fi — ?i,f,y hf (xi , y) , S m ? h2H i=1 f 2Fi y2Yf

where ? = (?i,f,y)i2[m],f 2Fi ,y2Yf and where ?i,f,y s are independent Rademacher random variables uniformly distributed over {?1}. The factor graph Rademacher complexity of H for a factor graph ? G ? b G is defined as the expectation: RG m (H) = ES?D m RS (H) . It can be shown that the empirical factor graph Rademacher complexity is concentrated around its mean (Lemma 8). The factor graph Rademacher complexity is a natural extension of the standard Rademacher complexity to vectorvalued hypothesis sets (with one

coordinate per factor in our case). For binary classification, the factor graph and standard Rademacher complexities coincide. Otherwise, the factor graph complexity can be upper bounded in terms of the standard one. As with the standard Rademacher complexity, the factor graph Rademacher complexity of a hypothesis set can be estimated from data in many cases. In some important cases, it also admits explicit upper bounds similar to those for the standard Rademacher complexity but with an additional dependence on the factor graph quantities. We will prove this for several families of functions which are commonly used in structured prediction (Theorem 2). 3.2

Generalization bounds

In this section, we present new margin bounds for structured prediction based on the factor graph Rademacher complexity of H. Our results hold both for the additive and the multiplicative empirical margin losses defined below: ? ? ? ? ? add ? 0 0 1 b RS,? (h) = E max L(y , y) ? h(x, y) h(x, y) (5) y 0 6=y (x,y)?S ? ? ? ? ? mult ? 0 0 1 bS,? R (h) = E max L(y , y) 1 [h(x, y) h(x, y)] . (6) ? 0 (x,y)?S

y 6=y

Here, ? (r) = min(M, max(0, r)) for all r, with M = maxy,y0 L(y, y 0). As we show in Section 5, badd (h) and R bmult (h) directly lead to many existing structured prediction convex upper bounds on R S,? S,? algorithms. The following is our general data-dependent margin bound for structured prediction. 4

Theorem 1. Fix ? ¿ 0. For any ¿ 0, with probability at least 1 of size m, the following holds for all h 2 H,

over the draw of a sample S

s p 1 badd (h) + 4 2 RG (H) + M log , R(h) ? R?add (h) ? R S,? m ? 2m s p log 1 4 2M G mult bS,? R(h) ? R?mult (h) ? R (h) + Rm (H) + M . ? 2m

The full proof of Theorem 1 is given in Appendix A. It is based on a new contraction lemma (Lemma 5) generalizing Talagrand?s lemma that can be of independent interest.2 We also present a more refined contraction lemma (Lemma 6) that can be used to improve the bounds of Theorem 1. Theorem 1 is the first data-dependent generalization guarantee for structured prediction with general loss functions, general hypothesis sets, and arbitrary factor graphs for both multiplicative and additive margins. We also present a version of this result with empirical complexities as Theorem 7 in the supplementary material. We will compare these guarantees to known special cases below. The margin bounds above p can be extended to hold uniformly over ? 2 (0, 1] at the price of an additional term of the form (log log2 ?2)/m in the bound, using known techniques (see for example [Mohri et al., 2012]). The hypothesis set used by convex structured prediction algorithms such as StructSVM [Tsochantaridis et al., 2005], Max-Margin Markov Networks (M3N) [Taskar et al., 2003] or Conditional Random Field (CRF) [Lafferty et al., 2001] is that of linearPfunctions. More precisely, let be a feature mapping from (X ? Y) to RN such that (x, y) = f 2F f (x, yf). For any p, define Hp as follows: Hp = {x 7! w ?

(x, y) : w 2 RN , kwkp ? ?p }.

b G (Hp) can be efficiently estimated using random sampling and solving

LP programs. Then, R m b G (Hp). To simplify our presentation, we will Moreover, one can obtain explicit upper bounds on R m consider the case p = 1, 2, but our results can be extended to arbitrary p 1 and, more generally, to arbitrary group norms. Theorem 2. For any sample S = (x1 , . . . , xm), the following upper bounds hold for the empirical factor graph complexity of H1 and H2 : q p P b G (H1) ? ?1 r1 s log(2N), b G (H2) ? ?2 r2 Pm P R R S S i=1 f 2Fi y2Yf —Fi —, m m where r1 = maxi,f,y k f (x , y)k 1 , r2 = Pim P Pmaxi,f,y k defined by s = maxj2[1,N] i=1 f 2Fi y2Yf —Fi —1

f (xi , y)k2

f,j (xi ,y)6=0

.

and where s is a sparsity factor

Plugging in these factor graph complexity upper bounds into Theorem 1 immediately yields explicit data-dependent structured prediction learning guarantees for linear hypotheses with general loss functions and arbitrary factor graphs (see Corollary 10). Observe that, in the worst case, the sparsity factor can be bounded as follows: s?

m X X X

i=1 f 2Fi y2Yf

—Fi — ?

m X i=1

—Fi —2 di ? m max —Fi —2 di , i

where di = max pf 2Fi —Yf —. Thus, the factor graph Rademacher complexities of linear hypotheses in H1 scale as O(log(N) maxi —Fi —2 di /m). An important observation is that —Fi — and di depend on the observed sample. This shows that the expected size of the factor graph is crucial for learning in this scenario. This should be contrasted with other existing structured prediction guarantees that we discuss below, which assume a fixed upper bound on the size of the factor graph. Note that our result shows that learning is possible even with an infinite set Y. To the best of our knowledge, this is the first learning guarantee for learning with infinitely many classes. 2

A result similar to Lemma 5 has also been recently proven independently in [Maurer, 2016].

5

Our learning guarantee for H1 can additionally benefit from the sparsity of the feature mapping and observed data. In particular, in many applications, f,j is a binary indicator function that is non-zero for a single (x, y) 2 X ? Yf . For instance, in NLP, P f,j may indicate an occurrence of a m 2 2 certain n-gram in the input xi and output y . In this case, s = i=1 —Fi — ? m maxi —Fi — and the p i complexity term is only in O(maxi —Fi — log(N)/m), where N may depend linearly on di . 3.3

Special cases and comparisons

Markov networks. For the pairwise Markov networks with a fixed number of substructures l studied by Taskar et al. [2003], our equivalent factor graph admits l nodes, —Fi — = l, and the maximum size of Yf is di = k 2 if each substructure of a pair can be assigned one of k classes. Thus, if we apply

Corollary 10 with Hamming distance as our loss function and divide the bound through by l, to normalize the loss to interval [0, 1] as in [Taskar et al., 2003], we obtain the following explicit form of our guarantee for an additive empirical margin loss, for all h 2 H2 : s r 2 log 1 4? r 2k 2 2 add bS,? R(h) ? R (h) + +3 . ? m 2m

This bound can be further improved by eliminating the dependency on k using an extension of our contraction Lemma 5 to k ? k1,2 (see p Lemma 6). The complexity term of Taskar et al. [2003] is e bounded by a quantity that varies as O(?22 q 2 r22 /m), where q is the maximal out-degree of a factor graph. Our bound has the same dependence on these key quantities, but with no logarithmic term in our case. Note that, unlike the result of Taskar et al. [2003], our bound also holds for general loss functions and different p-norm regularizers. Moreover, our result for a multiplicative empirical margin loss is new, even in this special case.

Multi-class classification. For standard (unstructured) multi-class classification, we have —Fi — = 1 and di = c, where c is the number of classes. In that case, for linear hypotheses with norm-2 p 2 regularization, the complexity term of our bound varies as O(?2 r2 c/? m) (Corollary 11). This improves upon the best known general margin bounds of Kuznetsov et al. [2014], who provide a guarantee that scales linearly with the number of classes instead. Moreover, in the special case where an individual wy is learned for each class y 2 [c], we retrieve the recent favorable bounds given by Lei et al. [2015], albeit with a somewhat simpler formulation. In that case, for any (x, y), all components of the feature vector (x, y) are zero, except (perhaps) for the N components corresponding to class y, where N is the dimension of wy . In view of that, for p example for a group-norm k ? k2,1 regularization, the complexity term of our bound varies as O(?r (log c)/?2 m), which matches the results of Lei et al. [2015] with a logarithmic dependency on c (ignoring some complex exponents of log c in their case). Additionally, note that unlike existing multi-class learning guarantees, our results hold for arbitrary loss functions. See Corollary 12 for further details. Our sparsity-based bounds can also be used to give bounds with logarithmic dependence on the number of classes when the features only take values in {0, 1}. Finally, using Lemma 6 instead of Lemma 5, the dependency on the number of classes can be further improved. We conclude this section by observing that, since our guarantees are expressed in terms of the average size of the factor graph over a given sample, this invites us to search for a hypothesis set H and predictor h 2 H such that the tradeoff between the empirical size of the factor graph and empirical error is optimal. In the next section, we will make use of the recently developed principle of Voted Risk Minimization (VRM) [Cortes et al., 2015] to reach this objective.

4

Voted Risk Minimization

In many structured prediction applications such as natural language processing and computer vision, one may wish to exploit very rich features. However, the use of rich families of hypotheses could lead to overfitting. In this section, we show that it may be possible to use rich families in conjunction with simpler

families, provided that fewer complex hypotheses are used (or that they are used with less mixture weight). We achieve this goal by deriving learning guarantees for ensembles of structured prediction rules that explicitly account for the differing complexities between families. This will motivate the algorithms that we present in Section 5. 6

Assume that we are given p families H1 , . . . , Hp of functions mapping from X ? Y to R. Define the PT ensemble family F = conv([pk=1 Hk), that is the family of functions f of the form f = t=1 ?t ht , where ? = (?1 , . . . , ?T) is in the simplex and where, for each t 2 [1, T], ht is in Hkt for some G G kt 2 [1, p]. We further assume that RG m (H1) ? Rm (H2) ? . . . ? Rm (Hp). As an example, the Hk s may be ordered by the size of the corresponding factor graphs. The main result of this section is a generalization of the VRM theory to the structured prediction badd (h) and setting. The learning guarantees that we present are in terms of upper bounds on R S,? bmult (h), which are defined as follows for all ? 0: R S,? ? ? ? ? ? add ? 0 0 1 bS,?,? R (h) = E max L(y , y) + ? h(x, y) h(x, y) (7) ? y 0 6=y (x,y)?S ? ? ? ?? mult ? 0 0 1 bS,?,? R (h) = E max L(y , y) 1 + ? [h(x, y) h(x, y)] . (8) ? 0 (x,y)?S

y 6=y

Here, ? can be interpreted as a margin term that acts in conjunction with ?. For simplicity, we assume in this section that —Y— = c ¡ +1.

Theorem 3. Fix ? ¿ 0. For any ¿ 0, with probability at least 1 over the draw of a sample S of size m, each of the following inequalities holds for all f 2 F: p T 4 2X add b R(f) RS,?,1 (f) ? ? t RG m (Hkt) + C(?, M, c, m, p), ? t=1 p T 4 2M X mult b R(f) RS,?,1 (f) ? ? t RG m (Hkt) + C(?, M, c, m, p), ? t=1 rl q m log 2 log p c2 ?2 m log p 2M 4 where C(?, M, c, m, p) = ? m + 3M ?2 log 4 log p m + 2m .

The proof of this theorem crucially depends on the theory we developed in Section 3 and is given in Appendix A. As with Theorem 1, we also present a version of this result with empirical complexities as Theorem 14 in the supplementary material. The explicit dependence of this bound on the parameter vector ? suggests that learning even with highly complex hypothesis sets could be possible so long as the complexity term, which is a weighted average of the factor graph complexities, is not too large. The theorem provides a quantitative way of determining the mixture weights that should be apportioned to each family. Furthermore, the dependency on the number of distinct feature map families Hk is very mild and therefore suggests that a large number of families can be used. These properties will be useful for motivating new algorithms for structured prediction.

5

Algorithms

In this section, we derive several algorithms for structured prediction based on the VRM principle discussed in Section 4. We first give general convex upper bounds (Section 5.1) on the structured prediction loss which recover as special cases the loss functions used in StructSVM [Tsochantaridis et al., 2005], Max-Margin Markov Networks (M3N) [Taskar et al., 2003], and Conditional Random Field (CRF) [Lafferty et al., 2001]. Next, we introduce a new algo-

rithm, Voted Conditional Random Field (VCRF) Section 5.2, with accompanying experiments as proof of concept. We also present another algorithm, Voted StructBoost (VStructBoost), in Appendix C. 5.1

General framework for convex surrogate losses

Given (x, y) 2 X ? Y, the mapping h 7! L(h(x), y) is typically not a convex function of h, which leads to computationally hard optimization problems. This motivates the use of convex surrogate losses. We first introduce a general formulation of surrogate losses for structured prediction problems. Lemma 4. For any u 2 R+ , let u : R ! R be an upper bound on v 7! u1v?0 . Then, the following upper bound holds for any h 2 H and (x, y) 2 X ? Y, L(h(x), y) ? max h(x, y 0)). (9) L(y 0 ,y) (h(x, y) 0 y 6=y

7

The proof is given in Appendix A. This result defines a general framework that enables us to straightforwardly recover many of the most common state-of-the-art structured prediction algorithms via suitable choices of u (v): (a) for u (v) = max(0, u(1 v)), the right-hand side of (9) coincides with the surrogate loss defining StructSVM [Tsochantaridis et al., 2005]; (b) for u (v) = max(0, u v), it coincides with the surrogate loss defining Max-Margin Markov Networks (M3N) [Taskar et al., 2003] when using for L the Hamming loss; and (c) for u (v) = log(1 + eu v), it coincides with the surrogate loss defining the Conditional Random Field (CRF) [Lafferty et al., 2001]. Moreover, alternative choices of u (v) can help define new algorithms. In particular, we will refer to the algorithm based on the surrogate loss defined by u (v) = ue v as StructBoost, in reference to the exponential loss used in AdaBoost. Another related alternative is based on the choice u (v) = eu v . See Appendix C, for further details on this algorithm. In fact, for each u (v) described above, the corresponding convex surrogate is an upper bound on either the multiplicative or additive margin loss introduced in Section 3. Therefore, each of these algorithms seeks a hypothesis that minimizes the generalization bounds presented in Section 3. To the best of our knowledge, this interpretation of these well-known structured prediction algorithms is also new. In what follows, we derive new structured prediction algorithms that minimize finer generalization bounds presented in Section 4. 5.2

Voted Conditional Random Field (VCRF)

We first consider the convex surrogate loss based on u (v) = log(1 + eu v), which corresponds to the loss defining CRF models. Using the monotonicity of the logarithm and upper bounding the maximum by a sum gives the following upper bound on the surrogate loss holds: ?X ? L(y,y 0) w?((x,y) (x,y 0)) L(y,y 0) w?((x,y) (x,y 0)) max log(1 + e) ? log e , 0 y 6=y

y 0 2Y

which, combined with VRM principle leads to the following optimization problem: ?X ? X p m 1 X min log eL(y,yi) w?((xi ,yi) (xi ,y)) + (rk +)kwk k1 , w m i=1 y2Y

(10)

k=1

p where rk = r1 —F (k)— log N . We refer to the learning algorithm based on the optimization problem (10) as VCRF. Note that for = 0, (10) coincides

with the objective function of L1 0 0 regularized CRF. Observe that we can also directly use maxy0 6=y log(1 + eL(y,y) w? (x,y,y)) or its P 0 0 upper bound y0 6=y log(1 + eL(y,y) w? (x,y,y)) as a convex surrogate. We can similarly derive an L2 -regularization formulation of the VCRF algorithm. In Appendix D, we describe efficient algorithms for solving the VCRF and VStructBoost optimization problems.

6

Experiments

In Appendix B, we corroborate our theory by reporting experimental results suggesting that the VCRF algorithm can outperform the CRF algorithm on a number of part-of-speech (POS) datasets.

7

Conclusion

We presented a general theoretical analysis of structured prediction. Our data-dependent margin guarantees for structured prediction can be used to guide the design of new algorithms or to derive guarantees for existing ones. Its explicit dependency on the properties of the factor graph and on feature sparsity can help shed new light on the role played by the graph and features in generalization. Our extension of the VRM theory to structured prediction provides a new analysis of generalization when using a very rich set of features, which is common in applications such as natural language processing and leads to new algorithms, VCRF and VStructBoost. Our experimental results for VCRF serve as a proof of concept and motivate more extensive empirical studies of these algorithms. Acknowledgments This work was partly funded by NSF CCF-1535987 and IIS-1618662 and NSF GRFP DGE-1342536. 8

2 References

G. H. Bakir, T. Hofmann, B. Sch?lkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan. Predicting Structured Data (Neural Information Processing). The MIT Press, 2007. K. Chang, A. Krishnamurthy, A. Agarwal, H. Daum? III, and J. Langford. Learning to search better than your teacher. In ICML, 2015. M. Collins. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In Proceedings of IWPT, 2001. C. Cortes, M. Mohri, and J. Weston. A General Regression Framework for Learning String-to-String Mappings. In Predicting Structured Data. MIT Press, 2007. C. Cortes, V. Kuznetsov, and M. Mohri. Ensemble methods for structured prediction. In ICML, 2014. C. Cortes, P. Goyal, V. Kuznetsov, and M. Mohri. Kernel extraction via voted risk minimization. JMLR, 2015. H. Daum? III, J. Langford, and D. Marcu. Search-based structured prediction. Machine Learning, 75(3): 297?325, 2009. J. R. Doppa, A. Fern, and P. Tadepalli. Structured prediction via output space search. JMLR, 15(1):1317?1350, 2014. D. Jurafsky and J. H. Martin. Speech and Language Processing (2nd Edition). Prentice-Hall, Inc., 2009. V. Kuznetsov, M. Mohri, and U. Syed. Multi-class deep boosting. In NIPS, 2014. J. Lafferty, A. McCallum, and F. Pereira. Condi-

tional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML, 2001. M. Lam, J. R. Doppa, S. Todorovic, and T. G. Dietterich. Hc-search for structured prediction in computer vision. In CVPR, 2015. Y. Lei, ?. D. Dogan, A. Binder, and M. Kloft. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. In NIPS, 2015. A. Lucchi, L. Yunpeng, and P. Fua. Learning for structured prediction using approximate subgradient descent with working sets. In CVPR, 2013. A. Maurer. A vector-contraction inequality for rademacher complexities. In ALT, 2016. D. McAllester. Generalization bounds and consistency for structured labeling. In Predicting Structured Data. MIT Press, 2007. M. Mohri, A. Rostamizadeh, and A. Talwalkar. Foundations of Machine Learning. The MIT Press, 2012. D. Nadeau and S. Sekine. A survey of named entity recognition and classification. Linguisticae Investigationes, 30(1):3?26, January 2007. S. Ross, G. J. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In AISTATS, 2011. B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In NIPS, 2003. I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. JMLR, 6:1453?1484, Dec. 2005. O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton. Grammar as a foreign language. In NIPS, 2015a. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In CVPR, 2015b. D. Zhang, L. Sun, and W. Li. A structured prediction approach for statistical machine translation. In IJCNLP, 2008.

9