

David McAllester

Toyota Technological Institute at Chicago

mcallester@tti-c.org

This chapter gives generalization bounds for structured output learning. We show that generalization bounds justify the use of Hamming distance in training algorithms independent of the choice of the loss function used to define generalization error. In particular, even when generalization error is measured by 0-1 loss the generalization bounds involve a Hamming distance. A natural training algorithm is to simply minimize the generalization bound over the concept parameters. Minimizing the generalization bound is consistent — as the amount of training data increases the performance of the algorithm approaches the minimal generalization error achievable over the parameter settings. Unfortunately, the generalization bound is not convex in concept parameters. We consider several different ways to make the generalization bound convex all of which are equivalent to classical hinge loss in the case of binary classification but none of which are consistent.

1.1 Introduction

Structured output classification can be viewed as a kind of decoding. We assume a probability distribution on pairs $\langle x, y \rangle$ where x is observable and y is latent. A decoder is a machine for predicting y given only x . In communication channels, as in structured labeling, one typically deals with cases where y is a structured signal.

In this chapter we will be concerned only with a kind of linear decoding. We assume a fixed mapping Φ from pairs to feature vectors, i.e., for any pair $\langle x, y \rangle$ we have $\Phi(x, y) \in \mathfrak{R}^d$. We will consider decoders of the following form where $w \in \mathfrak{R}^d$ is a weight vector.

$$f_w(x) = \operatorname{argmax}_y \Phi(x, y) \cdot w \tag{1.1}$$

The ultimate objective is to set the parameters w so as to minimize the expectation of $d(y, f_w(x))$ where d is a measure of distortion.

$$w^* = \operatorname{argmin}_w \mathbb{E}_{\langle x, y \rangle \sim D} [d(y, f_w(x))] \quad (1.2)$$

A popular alternative to (1.2) is logistic regression. In logistic regression the weight vector w is used to represent a probability distribution $P(y|x, w)$ defined as follows.

$$P(y|x, w) = \frac{1}{Z(x, w)} \exp(\Phi(x, y) \cdot w) \quad (1.3)$$

$$Z(x, w) = \sum_{\hat{y}} \exp(\Phi(x, \hat{y}) \cdot w) \quad (1.4)$$

Models of this form include Markov random fields (MRFs), probabilistic context free grammars (PCFGs), hidden Markov models (HMMs), conditional random fields (CRFs) (6), dynamic Bayes nets (5), and probabilistic relational models (PRMs) (4). In logistic regression the goal is to minimize expected log loss.

$$w^* = \operatorname{argmin}_w \mathbb{E}_{\langle x, y \rangle \sim D} \left[\log \frac{1}{P(y|x, w)} \right] \quad (1.5)$$

A significant advantage of logistic regression is that (1.5) is convex in w while (1.2) is not. However, the objective in (1.2) seems a more accurate reflection of the actual quantity of interest. The main question addressed in this chapter is how one should select the parameter vector w so as to approximate (1.2) given only a finite sample of training data drawn from D .

For binary classification, the case with $y \in \{-1, 1\}$, SVMs provide a popular approach to optimizing (1.2). But for the general case of structured decoding there are several different generalizations of binary SVMs. Here we give generalization bounds designed to provide insight into these alternatives.

Generalization bounds were given by Collins for 0-1 distortion (3) and a bound has been given by Taskar et al. (12) for the case of Hamming distance distortion. The use of Hamming distance produces a much tighter bound and seems to support the idea that Hamming distortion has advantages in practice. Here we show that the improvements in the generalization analysis achieved by Taskar et al. can be achieved for an arbitrary bounded distortion, including 0-1 distortion. Interestingly, Hamming distance still appears in the analysis, but not as a consequence of the choice of distortion. We also consider issues of asymptotic consistency. With more than two possible signals, SVM algorithms are not consistent — they fail to converge on the optimal decoder even in the limit of infinite training data. However, the training algorithm that sets w by minimizing the (nonconvex) generalization bound is consistent. This gives a trade-off between convexity and consistency in decoding with structured signals.

1.2 PAC-Bayesian Generalization Bounds

In the decoding problem considered here, the goal is to learn a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ where \mathcal{X} is a set of observable “codes” and \mathcal{Y} is a set of latent (unobserved) “signals”. Here we follow the approach given in (2) based on “parts”. In addition to the sets \mathcal{X} and \mathcal{Y} we assume a set \mathcal{P} of parts. In parsing we have that x is a string and y is a parse tree with yield x — the decoder takes as input a string and produces as output a parse tree. In parsing we have a stochastic (or weighted) grammar G . In parsing a part is just a production of G . Each pair $\langle x, y \rangle$ of a string x and a parse tree y is associated with a set of parts — the productions of G that appear in the parse tree y . Note that a given parse tree can use the same production more than once. Also note that for parsing with a finite grammar the set of parts is finite even though the spaces \mathcal{X} and \mathcal{Y} are infinite.

In general we assume sets \mathcal{X} , \mathcal{Y} and \mathcal{P} and we assume a function c such that for $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $p \in \mathcal{P}$ we have that $c(p, \langle x, y \rangle)$ is a nonnegative integer — $c(p, \langle x, y \rangle)$ gives the number of times that the part p appears in the pair $\langle x, y \rangle$. Furthermore, we assume a distribution on D on $\mathcal{X} \times \mathcal{Y}$ such that for any $x \in \mathcal{X}$ we have that the conditional distribution $P(y|x)$ has a countable support (a feasible set) which we denote by $\mathcal{Y}(x)$. We let $\mathcal{P}(x)$ denote the set of $p \in \mathcal{P}$ such that there exists $\hat{y} \in \mathcal{Y}(x)$ with $c(p, \langle x, \hat{y} \rangle) > 0$. Here we will assume that for any $x \in \mathcal{X}$ the sets $\mathcal{Y}(x)$ and $\mathcal{P}(x)$ are finite. For grammars we have that $c(p, \langle x, y \rangle)$ is the number of times the production p occurs in the parse tree y .

We consider the decoder f_w defined by (1.1) where $w, \Phi(x, y) \in \mathbb{R}^{|\mathcal{P}|}$ and where $\Phi(x, y)$ is defined as follows.

$$\Phi_p(x, y) = c(p, \langle x, y \rangle) \tag{1.6}$$

In the case of grammars we have that w and $\Phi(x, y)$ are both indexed by the set of productions of the grammar. A more general form for $\Phi(x, y)$, allowing for the use of kernels, is discussed in section 1.5. For any definition of $\Phi(x, y)$ we define the margin $m(x, y, \hat{y}, w)$ as follows.

$$m(x, y, \hat{y}, w) = \Phi(x, y) \cdot w - \Phi(x, \hat{y}) \cdot w \tag{1.7}$$

Intuitively, $m(x, y, \hat{y}, w)$ is the amount by which y is preferable to \hat{y} under the parameter setting w .

The PAC-Bayesian theorem governs the expected distortion of a stochastic decoder. The stochastic decoder first stochastically selects an alternative parameter vector w' then then returns $f_{w'}(x)$. It is possible to convert the bounds stated here for the loss of a stochastic decoder to a bound on the loss of the deterministic decoder f_w . However, this conversion seems to provides no additional insight. For any weight vector w we let $Q(w)$ be a distribution on weight vectors whose precise

definition is given in section 1.6. We define the expected distortion of $Q(w)$ as follows.

$$L(Q(w), D) = \mathbb{E}_{\langle x, y \rangle \sim D, w' \sim Q(w)} [d(y, f_{w'}(x))] \quad (1.8)$$

For simplicity we assume that there exist finite values r , s and ℓ satisfying the following conditions for all $x \in \mathcal{X}$ and $\hat{y} \in \mathcal{Y}(x)$.

$$|\mathcal{Y}(x)| \leq r \quad (1.9)$$

$$\sum_{p \in \mathcal{P}(x)} c(p, \langle x, \hat{y} \rangle) \leq s \quad (1.10)$$

$$|\mathcal{P}(x)| \leq \ell \quad (1.11)$$

In parsing we have that finite values r , s , and ℓ exist provided that we bound the length n of the string x . In this case r is exponential in n while s is $O(n)$ and ℓ is $O(1)$ (the number of productions of the grammar). In lexicalized grammars bounds r , s and ℓ can be given in terms of the length of the input string independent of the size of the lexicon.

Throughout the rest of this paper we assume that $0 \leq d(y, \hat{y}) \leq 1$ for all $y, \hat{y} \in \mathcal{Y}(x)$ with $x \in \mathcal{X}$. We also assume a sequence $S = \langle x_1, y_1 \rangle, \dots, \langle x_m, y_m \rangle$ of m labeled training pairs has been drawn IID from D . The following theorem is similar to that proved by Collins but generalized to handle arbitrary bounded distortion and modified so as to be consistent. More specifically, if we set w^* so as to minimize the right hand side of (1.12) then, in the limit of infinite training data, we have that w^* minimizes generalization loss. This is discussed in more detail in section 1.4. In the theorem $I[\Phi]$ denotes the indicator function where $I[\Phi] = 1$ if Φ is true and 0 otherwise.

Theorem 1 *With probability at least $1 - \delta$ over the draw of the training data S of m pairs we have that the following holds simultaneously for all weight vectors w .*

$$L(Q(w), D) \leq \frac{\mathcal{L}_1(w, S)}{m} + \frac{\|w\|^2}{m} + \sqrt{\frac{2s^2\|w\|^2 \ln\left(\frac{rm}{\|w\|^2}\right) + \ln\left(\frac{m}{\delta}\right)}{(m-1)}} \quad (1.12)$$

$$\mathcal{L}_1(w, S) = \sum_{i=1}^m \max_{\hat{y} \in \mathcal{Y}(x_i)} d(y_i, \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq 1] \quad (1.13)$$

Intuitively, $f_w(x_i)$ is uncertain and any value \hat{y} satisfying $m(x_i, f_w(x_i), \hat{y}, w) \leq 1$ should be considered as a possible value of $f_w(x_i)$. The quantity $\mathcal{L}_1(w, S)$ is the worst case distortion over the signals \hat{y} which are considered to be possible values of $f_w(x_i)$. If all possible values of $f_w(x_i)$ are similar to y_i (as measured by the distortion function) then $\mathcal{L}_1(w, S)$ will be low. Theorem 1 should be contrasted with the following which refines the bound of Taskar et al. by handling arbitrary bounded distortion and modified so as to be consistent.

Theorem 2 *With probability at least $1 - \delta$ over the choice of the training data we have that the following holds simultaneously for all weight vectors w .*

$$L(Q(w), D) \leq \frac{\mathcal{L}_H(w, S)}{m} + \frac{\|w\|^2}{m} + \sqrt{\frac{\|w\|^2 \ln\left(\frac{2\ell m}{\|w\|^2}\right) + \ln\left(\frac{m}{\delta}\right)}{2(m-1)}} \quad (1.14)$$

$$\mathcal{L}_H(w, S) = \sum_{i=1}^m \max_{\hat{y} \in \mathcal{Y}(x_i)} d(y_i, \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq H(x_i, f_w(x_i), \hat{y})] \quad (1.15)$$

$$H(x, y, \hat{y}) = \sum_{p \in \mathcal{P}(x)} |c(p, \langle x, \hat{y} \rangle) - c(p, \langle x, y \rangle)| \quad (1.16)$$

Like $\mathcal{L}_1(w, S)$, the training loss $\mathcal{L}_H(w, S)$ can be viewed as the worst case distortion on the training set over the training labels that are considered possible values of $f_w(x_i)$. However, the criterion for being considered to be a possible value for $f_w(x_i)$ involves a Hamming distance.

The proof of both theorems is given in section 1.6. The main feature of the proof of theorem 1 is a union bound over the elements of $\mathcal{Y}(x_i)$ leading to the appearance of r in the bound. The bound is also influenced by the fact that $\|\Phi(x, y)\|^2$ can be as large as s^2 . The proof of theorem 2 replaces the union bound over $\mathcal{Y}(x)$ by a union bound over $\mathcal{P}(x)$ which is typically exponentially smaller.

At first theorem 1 and 2 may appear incomparable. However, theorem 2 dominates theorem 1 when $\ell \ll r$. To establish this one must make the two margin requirements comparable. Margin requirements can be changed by rescaling the weight vector. It is well known that in support vector machines one can either work with unit norm weight vectors and bounds involving the margin, or work with unit margin and bounds involving the norm of the weight vector. To compare theorems 1 and 2 we insert $w/(2s)$ for w in theorem 1 to get the following equivalent statement.

$$L\left(Q\left(\frac{w}{2s}\right), D\right) \leq \frac{\mathcal{L}_{2s}(w, S)}{m} + \frac{\|w\|^2}{4s^2 m} + \sqrt{\frac{\|w\|^2 \ln\left(\frac{4s^2 r m}{\|w\|^2}\right) + \ln\left(\frac{m}{\delta}\right)}{2(m-1)}} \quad (1.17)$$

$$\mathcal{L}_{2s}(w, S) = \sum_{i=1}^m \max_{\hat{y} \in \mathcal{Y}(x_i)} d(y_i, \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq 2s] \quad (1.18)$$

We now compare (1.17) with (1.14). We ignore the fact that the posteriors $Q(w)$ and $Q(w/2s)$ are different. Showing that the right hand side of (1.14) dominates the right hand of (1.17) shows that theorem 2 can provide a better guarantee than theorem 1 even if that better guarantee is for a different classifier. We can also justify ignoring the difference between $Q(w)$ and $Q(w/2s)$ with the claim that variants of these bounds can be proved for deterministic classifiers and the deterministic classifiers f_w and $f_{w/2s}$ are the same. To compare the right hand sides of (1.17) and (1.14) we first note that $H(x_i, f_w(x_i), \hat{y}) \leq 2s$ and therefore $\mathcal{L}_H(w, S) \leq \mathcal{L}_{2s}(w, S)$. Furthermore, for structured problems we have that r is exponentially larger than ℓ and hence the regularization term in (1.17) is larger than the regularization term in (1.14).

1.3 Hinge Loss

Support vector machines (SVMs) provide a popular alternative to logistic regression for binary classification. In this section we consider generalizations of SVMs to structured decoding. SVMs involve the optimization of hinge loss. When discussing hinge loss and its relationship to generalization bounds the following notation for the step function and the ramp function will be useful.

$$\begin{aligned}(x)^+ &= I[x \geq 0] \\ (x)_+ &= \min(0, x)\end{aligned}$$

In the case of binary classification we have $y \in \{1, -1\}$. In this case we have that (1.1) can be written as follows.

$$f_w(x) = \operatorname{argmax}_{y \in \{-1, 1\}} \Phi(x, y) \cdot w = \operatorname{sign}(\Phi(x) \cdot w) \quad (1.19)$$

where

$$\Phi(x, 1) = -\Phi(x, -1) = \frac{1}{2}\Phi(x) \quad (1.20)$$

A support vector machine selects w so as to minimize the following regularized hinge loss objective function where $y_i(\Phi(x_i) \cdot w)$ is called the margin and $(1 - m)_+$ is called the hinge loss of margin m .

$$w^* = \operatorname{argmin}_w \sum_i (1 - y_i(\Phi(x_i) \cdot w))_+ + \lambda \|w\|^2 \quad (1.21)$$

Collins (3) considered structured SVMs using a multiclass hinge loss.

$$w^* = \operatorname{argmin}_w \sum_i \max_{\hat{y} \neq y_i} (1 - m(x_i, y_i, \hat{y}, w))_+ + \lambda \|w\|^2 \quad (1.22)$$

Altun and Hoffman (1) proposed the following.

$$w^* = \operatorname{argmin}_w \sum_i \max_{\hat{y}} d(y_i, \hat{y}) (1 - m(x_i, y_i, \hat{y}, w))_+ + \lambda \|w\|^2 \quad (1.23)$$

Taskar Guestrin and Koller, (12), proposed the following.

$$w^* = \operatorname{argmin}_w \sum_i \max_{\hat{y}} (H(x_i, y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w))_+ + \lambda \|w\|^2 \quad (1.24)$$

The optimizations (1.22), (1.23), and (1.24) all reduce to (1.21) in the case of binary classification. The fact that theorem 2 dominates theorem 1 suggests that (1.24) is preferable to (1.22) or (1.23). But the precise relationship between the

generalization bounds and the various notions of hinge loss is subtle. Theorems 1 and 2 directly motivate the following.

$$w^* = \operatorname{argmin}_w \sum_i \max_{\hat{y}} d(y_i, \hat{y}) (1 - m(x_i, f_w(x_i), \hat{y}, w))^+ + \lambda \|w\|^2 \quad (1.25)$$

$$w^* = \operatorname{argmin}_w \sum_i \max_{\hat{y}} d(y_i, \hat{y}) \left(\begin{array}{c} H(x_i, f_w(x_i), \hat{y}) \\ -m(x_i, f_w(x_i), \hat{y}, w) \end{array} \right)^+ + \lambda \|w\|^2 \quad (1.26)$$

The optimization problems given by (1.25) and (1.26) are not convex in w . As a first step in approximating these by convex functions we can replace the step functions by ramps. This replacement yields the following.

$$w^* = \operatorname{argmin}_w \sum_i \max_{\hat{y}} d(y_i, \hat{y}) (1 - m(x_i, f_w(x_i), \hat{y}, w))_+ + \lambda \|w\|^2 \quad (1.27)$$

$$w^* = \operatorname{argmin}_w \sum_i \max_{\hat{y}} d(y_i, \hat{y}) \left(\begin{array}{c} H(x_i, f_w(x_i), \hat{y}) \\ -m(x_i, f_w(x_i), \hat{y}, w) \end{array} \right)_+ + \lambda \|w\|^2 \quad (1.28)$$

But (1.27) and (1.28) are still not convex. To see this consider the case of binary classification under 0-1 distortion. Because $d(y_i, \hat{y}) = 0$ for $\hat{y} = y_i$ we need consider only the case where $\hat{y} \neq y_i$. If $f_w(x_i) = y_i$ then $(1 - m(x_i, f_w(x_i), \hat{y}, w))_+$ equals the binary hinge loss $(1 - y_i(\Phi(x) \cdot w))_+$. But if $f_w(x_i) \neq y_i$ — the case where the classical margin is less than zero — then $(1 - m(x_i, f_w(x_i), \hat{y}, w))_+ = 1$. So in the binary case $(1 - m(x_i, f_w(x_i), \hat{y}, w))_+$ is a continuous piecewise linear non-convex function which equals the hinge loss for positive margin but equals the constant 1 for negative margin. A second step in making this convex is to replace $f_w(x_i)$ by the constant y_i . With this replacement (1.27) becomes (1.23) and (1.28) becomes the following.

$$w^* = \operatorname{argmin}_w \sum_i \max_{\hat{y}} d(y_i, \hat{y}) (H(x_i, y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w))_+ + \lambda \|w\|^2 \quad (1.29)$$

It is interesting to note that (1.29) also reduce to (1.21) in the case of binary classification. It is also interesting to note that replacing $f_w(x_i)$ by y_i in theorems 1 and 2 strictly weakens the bounds and causes them to be inconsistent.

1.4 Consistency

Consistency is an important criterion for generalization bounds. More specifically, a bound is consistent if, in the limit of infinite data, the minimum of the bound (over the parameter vector w) approaches the minimum distortion possible over the allowed parameter space. The bounds in theorems 1 and 2 are both consistent in this sense. We give a quick sketch of a proof for this in the finite dimensional case where we have $w \in \mathfrak{R}^d$. We consider only theorem 1, the argument for theorem 2 is similar.

Since the unit sphere in \mathfrak{R}^d is compact, there must exist a vector w^* on the unit sphere minimizing generalization loss.

$$w^* = \operatorname{argmin}_{w: \|w\|=1} \mathcal{E}(w)$$

$$\mathcal{E}(w) = \mathbb{E}_{\langle x, y \rangle \sim D} [d(y, f_w(x))]$$

All vectors in the same direction as w^* yield the same classification function and hence the same expected distortion. We define w_m^* to be the vector $m^{1/3}w^*$ which is in the same direction as w^* but has length $m^{1/3}$. For a sample of size m we consider the value of the generalization bound for the vector w_m^* . Note that as $m \rightarrow \infty$ we have that $\|w_m^*\| \rightarrow \infty$ but the regularization term for w_m^* goes to zero. Furthermore, for vectors of sufficiently large norm, the only \hat{y} satisfying $m(x_i, f_w(x_i), \hat{y}) < 1$ is $f_w(x_i)$. This means that for vectors of sufficiently large norm we have that $\mathcal{L}_1(w, S)/m$ is essentially equivalent to the sample average of $d(x_i, y_i, f_w(x_i))$. Putting these two observations together we get that as $m \rightarrow \infty$ we have that the generalization bound for w_m^* must approach $\mathcal{E}(w^*)$. Hence, as $m \rightarrow \infty$ the minimum of the generalization bound is at most $\mathcal{E}(w^*)$. The algorithm is guaranteed (with high probability) to perform as well as the minimum of the generalization bound.

It is well known that the use of hinge loss in multiclass classification results in inconsistent algorithms (8). The various forms of convex hinge loss discussed in section 1.3 all fail to be consistent. It is possible to construct consistent forms of hinge loss for nonparametric, i.e., infinite feature dimension, multiclass classification (8). However, neither the convergence rates nor the practicality of these constructions has been established for the case of learning decoders.

To show inconsistency of the Hinge losses considered in section 1.3, suppose that \mathcal{X} contains only a single element x and that \mathcal{Y} is a finite set y_1, y_2, \dots, y_k , and that we are using 0–1 distortion. Further assume that there is an independent weight for each y_i . In other words, $d = k$ and $\Phi_j(x, y_i)$ is 1 if $i = j$ and zero otherwise so that $\Phi(x, y_i) \cdot w = w_i$. In this case all four of the hinge loss rules (1.22), (1.23), (1.24), and (1.29) are the same. We will work with the simplest form, (1.22). Assume that $\lambda = 0$ so that we simply want to minimize the hinge loss independent of $\|w\|^2$. In the limit of an infinitely large sample we have that the hinge loss term dominates the regularization term. Furthermore, suppose that for each y_i we have that the probability of the pair $\langle x, y_i \rangle$ is less than 1/2 (note that this cannot happen in binary classification). Define the margin m_i as follows.

$$m_i = w_i - \max_{j \neq i} w_j$$

In the limit of infinite training data we have the following.

$$w^* = \operatorname{argmin}_w \sum_i p_i (1 - m_i)_+ \tag{1.30}$$

We will show that in this case the optimal value is achieved when all weights are the same so that $m_i = 0$ for all i . To see this consider any uniform vector w . Since the objective function in (1.30) is convex it suffices to show that any variation in w fails to improve the objective function. As an example of a simple variation suppose that we increase the weight of the component w_i corresponding to the choice minimizing expected distortion. As we increase w_i we increase m_i but we decrease each m_j for $j \neq i$ by the same amount. Given that $p_i < 1/2$, the net effect is an increase in the objective function. To prove that no variation from uniform improves the objective, let $\Delta \in \mathfrak{R}^k$ be a vector and consider $w' = w + \epsilon\Delta$. We then have the following.

$$\frac{\partial m_i}{\partial \epsilon} = \Delta_i - \max_{j \neq i} \Delta_j$$

To show that is a non-improving variation it suffices to show the following.

$$\sum_i p_i (\Delta_i - \max_{j \neq i} \Delta_j) \leq 0$$

But this is equivalent to the following

$$\sum_i p_i \max_{j \neq i} \Delta_j \geq \sum_i p_i \Delta_i$$

This can be derived as follows where i^* is $\operatorname{argmax}_i \Delta_i$ and j^* is $\operatorname{argmax}_{j \neq i^*} \Delta_j$ and the last line follows from the assumption that $p_i < 1/2$.

$$\begin{aligned} \sum_i p_i \max_{j \neq i} \Delta_j &= p_{i^*} \Delta_{j^*} + \sum_{j \neq i^*} p_j \Delta_{i^*} \\ &= p_{i^*} \Delta_{j^*} + \sum_{j \neq i^*} p_j (\Delta_j + (\Delta_{i^*} - \Delta_j)) \\ &\geq p_{i^*} \Delta_{j^*} + \left(\sum_{j \neq i^*} p_j \Delta_j \right) + (1 - p_{i^*}) (\Delta_{i^*} - \Delta_{j^*}) \\ &= p_{i^*} \Delta_{i^*} + p_{i^*} (\Delta_{j^*} - \Delta_{i^*}) + \left(\sum_{j \neq i^*} p_j \Delta_j \right) + (1 - p_{i^*}) (\Delta_{i^*} - \Delta_{j^*}) \\ &= \sum_i p_i \Delta_i + (\Delta_{i^*} - \Delta_{j^*}) (1 - 2p_{i^*}) \\ &\geq \sum_i p_i \Delta_i \end{aligned}$$

Now we argue for the consistency of (1.25) and (1.26). If we hold λ/n fixed and let the sample size go to infinity then (1.25) and (1.26) become the following where $\lambda' = \lambda/n$

$$w^* = \operatorname{argmin}_w \mathbb{E}_{\langle x, y \rangle \sim D} \left[\max_{\hat{y}} d(y, \hat{y}) (1 - m(x, f_w(x), \hat{y}, w))^+ \right] + \lambda' \|w\|^2 \quad (1.31)$$

$$w^* = \operatorname{argmin}_w \mathbb{E}_{\langle x, y \rangle \sim D} \left[\max_{\hat{y}} d(y, \hat{y}) \left(\begin{array}{c} H(x, f_w(x), \hat{y}) \\ -m(x, f_w(x), \hat{y}, w) \end{array} \right)^+ \right] + \lambda' \|w\|^2 \quad (1.32)$$

Now we consider the limit of (1.31) and (1.32) as $\lambda' \rightarrow 0$. Intuitively this limit gives the following.

$$w^* = \operatorname{argmin}_w \mathbb{E}_{\langle x, y \rangle \sim D} \left[\max_{\hat{y}} d(y, \hat{y}) (1 - m(x, f_w(x), \hat{y}, w))^+ \right] \quad (1.33)$$

$$w^* = \operatorname{argmin}_w \mathbb{E}_{\langle x, y \rangle \sim D} \left[\max_{\hat{y}} d(y, \hat{y}) (H(x, f_w(x), \hat{y}) - m(x, f_w(x), \hat{y}, w))^+ \right] \quad (1.34)$$

The optimizations (1.33) and (1.34) yield very large vectors which drive any nonzero margin to be arbitrarily large. The direction of the optimal vector for both (1.33) and (1.34) is then given by the following.

$$w^* = \operatorname{argmin}_w \mathbb{E}_{\langle x, y \rangle \sim D} [d(y, f_w(x))] \quad (1.35)$$

To convert this argument into a formal proof one needs to give an explicit schedule for λ as a function of sample size and show that this schedule corresponds to taking m to infinity holding λ' constant and then taking λ' to zero. It should suffice to consider the “schedule” obtained by optimizing λ with hold-out data.

1.5 A Generalization of Theorem 2

We define two steps of increasing generality — a generalization to allow for kernels and a second generalization that generalizes the notion of part. The first generalization replaces (1.6) by the following where Ψ is a feature map on parts.

$$\Phi(x, y) = \sum_{p \in \mathcal{P}(x)} c(p, \langle x, y \rangle) \Psi(p) \quad (1.36)$$

This generalization is important when the parts themselves contain vectorial data. For example, in speech recognition the observable state in an HMM is often taken to be an acoustic feature vector. In (1.36) we have $\Psi(p) \in \mathfrak{R}^d$ where we allow $d = \infty$ with the understanding that \mathfrak{R}^∞ is the vector space of square summable infinite sequences. For the $d = \infty$ case it is usually more convenient to work in a

reproducing kernel Hilbert space (RKHS) defined by a kernel K in which case the decoder specified by weight function g is defined as follows.

$$f_g(x) = \operatorname{argmax}_{\hat{y} \in \mathcal{Y}(x)} \sum_{p \in \mathcal{P}(x)} c(p, \langle x, \hat{y} \rangle) g(p) \quad (1.37)$$

Formulation (1.37) is equivalent to (1.36) with $d = \infty$ and we will work only with (1.36).

We now generalize the notion of part. As before we now assume sets \mathcal{X} and \mathcal{Y} with a distribution D on $\mathcal{X} \times \mathcal{Y}$ with the property that for all $x \in \mathcal{X}$ the marginal $P(\cdot|x)$ has a countable support $\mathcal{Y}(x)$. We also assume a feature map Φ with $\Phi(x, y) \in \mathfrak{R}^d$ where, as above, we allow $d = \infty$. But rather than assume parts, we assume that for each $x \in \mathcal{X}$ we are given a set of $\ell(x)$ independent vectors $B(x)$ such that for all $\hat{y} \in \mathcal{Y}(x)$ we have that $\Phi(x, \hat{y})$ is in the linear span of $B(x)$. The vectors $\Psi(p)$ in (1.36) form such a basis. We let $\Psi_i(x)$ denote the i th vector in the basis $B(x)$. We can then generalize (1.36) to the following.

$$\Phi(x, y) = \sum_{i=1}^{\ell(x)} \gamma_i(x, y) \Psi_i(x) \quad (1.38)$$

The difference between (1.36) and (1.38) is actually quite minor. In (1.38) we have that $\gamma_i(x, y)$ is any real number while in (1.36) we have that $c(p, \langle x, y \rangle)$ must be a count — a non-negative integer. We now assume two quantities ℓ and R such that for all $x \in \mathcal{X}$ we have the following.

$$\ell(x) \leq \ell \quad (1.39)$$

$$\|\Psi_i(x)\| \leq R \quad (1.40)$$

We now state the generalization of theorem 2.

Theorem 3 *With probability at least $1 - \delta$ over the choice of the training data we have that the following holds simultaneously for all weight vectors w .*

$$L(Q(w), D) \leq \frac{\mathcal{L}_H(w, S)}{m} + \frac{R^2 \|w\|^2}{m} + \sqrt{\frac{R^2 \|w\|^2 \ln \left(\frac{2\ell m}{R^2 \|w\|^2} \right) + \ln \left(\frac{m}{\delta} \right)}{2(m-1)}} \quad (1.41)$$

$$\mathcal{L}_H(w, S) = \sum_{i=1}^m \max_{\hat{y} \in \mathcal{Y}(x_i)} d(y_i, \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq H(x_i, f_w(x_i), \hat{y})] \quad (1.42)$$

$$H(x, y, \hat{y}) = \sum_{i=1}^{\ell(x)} |\gamma_i(x, \hat{y}) - \gamma_i(x, y)| \quad (1.43)$$

The proof of this more general theorem is a straightforward generalization of the proof of theorem 2 and is not given here.

1.6 Proofs of Theorems 1 and 2

All our proofs use the PAC-Bayesian theorem (9; 11; 7; 10).

Lemma 4 (PAC-Bayesian Theorem) *For sets \mathcal{X} and \mathcal{Y} , any probability distribution D on $\mathcal{X} \times \mathcal{Y}$, any distortion function d on $\mathcal{Y} \times \mathcal{Y}$ with $0 \leq d(y, \hat{y}) \leq 1$, any decoder $f_w : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by parameter vector w , and any prior probability density P on the parameters w , we have that with probability at least $1 - \delta$ over the drawn of a sample $S = \langle \langle x_1, y_1 \rangle, \dots, \langle x_m, y_m \rangle \rangle$ from distribution D that the following holds simultaneously for all densities Q on parameters.*

$$L(Q, D) \leq L(Q, S) + \sqrt{\frac{KL(Q, P) + \ln \frac{m}{\delta}}{2(m-1)}} \quad (1.44)$$

where

$$L(Q, D) = \mathbb{E}_{\langle x, y \rangle \sim D, w \sim Q} [d(y, f_w(x))] \quad (1.45)$$

$$L(Q, S) = \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{w \sim Q} [d(y_i, f_w(x_i))] \quad (1.46)$$

Quadratic regularization corresponds to a Gaussian prior. We consider the $d = \infty$ case of a Gaussian process prior as the limit of the finite d case as d increases without bound. More specifically, we take the ‘‘prior’’ density to be a unit-variance isotropic Gaussian on weight vectors defined as follows.

$$p(w) = \frac{1}{Z} \exp\left(-\frac{\|w\|^2}{2}\right)$$

Theorems 1 and 2 govern the distortion of a stochastic decoder that stochastically draws w' from a distribution $Q(w)$. We now define the density $Q(w)$ as follows.

$$q(w' | w) = \frac{1}{Z} \exp\left(-\frac{1}{2} \|(w' - \alpha w)\|^2\right) \quad (1.47)$$

Here α is a scalar multiple which will be optimized later. The distribution $Q(w)$ is a unit-variance Gaussian about centered at αw . If α is very large then the vast majority of vectors drawn from $Q(w)$ will be in essentially the same direction as w . So by tuning α we can tune the degree to which $Q(w)$ is concentrated on vectors pointing in the same direction as w . The KL divergence from Q to P can be solved analytically as follows.

$$KL(Q(w) || P) = \frac{\alpha^2 \|w\|^2}{2} \quad (1.48)$$

To apply the PAC-Bayesian theorem it remains only to analyze the training loss $L(Q(w), S)$. In analyzing the training loss we can consider each training point $\langle x_i, y_i \rangle$ independently.

$$L(Q, S) = \frac{1}{m} \sum_{i=1}^m L_i \quad (1.49)$$

$$L_i = \mathbb{E}_{w' \sim Q(w)} [d(y_i, f_{w'}(x_i))] \quad (1.50)$$

In analyzing L_i we have that $f_{w'}(x_i)$ is a random variable based on the random draw of w' . The difference between theorem 1 and theorem 2 involves a different way of analyzing the random variable $f_{w'}(x_i)$. For theorem 1 we use the following lemma.

Lemma 5 For s and r as defined at the start of section 1.2 and α defined by

$$\alpha = s \sqrt{8 \ln \left(\frac{rm}{\|w\|^2} \right)}$$

we have that with probability at least $1 - \frac{\|w\|^2}{m}$ over the selection of w' the following holds.

$$f_{w'}(x_i) \in \{\hat{y} : m(x_i, f_w(x_i), \hat{y}, w) \leq 1\}$$

Proof Let \hat{y}_i abbreviate $f_w(x_i)$. We first note that by a union bound over the elements of $\mathcal{Y}(x_i)$ it suffices to show that for any given \hat{y} with $m(x_i, \hat{y}_i, \hat{y}, w) \geq 1$, the probability that $f_{w'}(x_i) = \hat{y}$ is at most $\|w\|^2/(rm)$. Consider a fixed $\hat{y} \in \mathcal{Y}(x_i)$ with $m(x_i, \hat{y}_i, \hat{y}, w) \geq 1$. We analyze the probability that the choice of w' overcomes the margin and causes \hat{y} to have a better score than \hat{y}_i . We first note the following for any vector $\Psi \in R^d$ with $\|\Psi\| = 1$ and any $\epsilon \geq 0$.

$$\mathbb{P}_{w' \sim Q(w)} [(\alpha w - w') \cdot \Psi \geq \epsilon] \leq \exp\left(\frac{-\epsilon^2}{2}\right) \quad (1.51)$$

For $\Delta(x_i, \hat{y}_i, \hat{y}) = \Phi(x_i, \hat{y}_i) - \Phi(x_i, \hat{y})$ we then have the following.

$$m(x_i, \hat{y}_i, \hat{y}, w) = \Delta(x_i, \hat{y}_i, \hat{y}) \cdot w$$

$$\|\Delta(x_i, \hat{y}_i, \hat{y})\| \leq 2s$$

Inserting $\Delta(x_i, \hat{y}_i, \hat{y})/\|\Delta(x_i, \hat{y}_i, \hat{y})\|$ into (1.51) yields the following.

$$\begin{aligned} \mathbb{P}_{w' \sim Q(w)} [m(x_i, \hat{y}_i, \hat{y}, w') \leq \alpha m(x_i, \hat{y}_i, \hat{y}, w) - \epsilon \|\Delta(x_i, \hat{y}_i, \hat{y})\|] &\leq \exp\left(\frac{-\epsilon^2}{2}\right) \\ \mathbb{P}_{w' \sim Q(w)} [m(x_i, \hat{y}_i, \hat{y}, w') \leq \alpha - \epsilon \|\Delta(x_i, \hat{y}_i, \hat{y})\|] &\leq \exp\left(\frac{-\epsilon^2}{2}\right) \end{aligned}$$

Setting ϵ equal to $\alpha/||\Delta(x_i, \hat{y}_i, \hat{y})||$ we get the following.

$$\begin{aligned} \mathbb{P}_{w' \sim Q(w)} [m(x_i, \hat{y}_i, \hat{y}, w') \leq 0] &\leq \exp\left(\frac{-\alpha^2}{2||\Delta(x_i, \hat{y}_i, \hat{y})||^2}\right) \\ \mathbb{P}_{w' \sim Q(w)} [f_{w'}(x_i) = \hat{y}] &\leq \exp\left(\frac{-\alpha^2}{2||\Delta(x_i, \hat{y}_i, \hat{y})||^2}\right) \\ &\leq \exp\left(\frac{-\alpha^2}{8s^2}\right) \end{aligned}$$

Setting α as in the statement of the lemma finishes the proof. \blacksquare

Theorem 1 now follows from the PAC-Bayesian theorem, (1.48), and lemma 5. Theorem 2 follows from the PAC-Bayesian theorem, (1.48), and the following lemma which is similar in form to lemma 5.

Lemma 6 For s and ℓ as defined at the start of section 1.2 and α defined by

$$\alpha = \sqrt{2 \ln \left(\frac{2\ell m}{||w||^2} \right)}$$

we have that with probability at least $1 - \frac{||w||^2}{m}$ over the selection of w' the following holds.

$$f_{w'}(x_i) \in \{\hat{y} : m(x_i, f_w(x_i), \hat{y}, w) \leq H(x_i, f_w(x_i), \hat{y}, w)\} \quad (1.52)$$

Proof Again let \hat{y}_i denote $f_w(x_i)$. First we note that for any $p \in \mathcal{P}(x_i)$ we have the following.

$$\mathbb{P}_{w' \sim Q(w)} [|w'_p - \alpha w_p| \geq \epsilon] \leq 2 \exp\left(\frac{-\epsilon^2}{2}\right) \quad (1.53)$$

Setting ϵ to α , for the above value of α , gives the following.

$$\mathbb{P}_{w' \sim Q(w)} [|w'_p - \alpha w_p| \geq \alpha] \leq \frac{||w||^2}{\ell m} \quad (1.54)$$

Now taking a union bound over the elements of $\mathcal{P}(x_i)$ we get that with probability $1 - \frac{||w||^2}{m}$ the following holds simultaneously for all $p \in \mathcal{P}(x_i)$.

$$|w'_p - \alpha w_p| \leq \alpha \quad (1.55)$$

Now assume that (1.55) holds for all $p \in \mathcal{P}(x_i)$. Consider \hat{y} such that $m(x_i, \hat{y}_i, \hat{y}, w) > H(x_i, \hat{y}_i, \hat{y})$. Let $\Delta(x_i, \hat{y}_i, \hat{y})$ denote $\Phi(x_i, \hat{y}_i) - \Phi(x_i, \hat{y})$. We now have the following.

$$m(x_i, \hat{y}_i, \hat{y}, w') \tag{1.56}$$

$$= m(x_i, \hat{y}_i, \hat{y}, \alpha w + (w' - \alpha w)) \tag{1.57}$$

$$= \alpha m(x_i, \hat{y}_i, \hat{y}, w) - \Delta(x_i, \hat{y}_i, \hat{y}) \cdot (\alpha w - w') \tag{1.58}$$

$$> \alpha H(x_i, \hat{y}_i, \hat{y}) - \Delta(x_i, \hat{y}_i, \hat{y}) \cdot (\alpha w - w') \tag{1.59}$$

$$\geq \alpha H(x_i, \hat{y}_i, \hat{y}) - \sum_{p \in \mathcal{P}(x_i)} (c(p, \langle x_i, \hat{y}_i \rangle) - c(p, \langle x_i, \hat{y} \rangle)) |\alpha w_p - w'_p| \tag{1.60}$$

$$\geq \alpha H(x_i, \hat{y}_i, \hat{y}) - \sum_{p \in \mathcal{P}(x_i)} |c(p, \langle x_i, \hat{y}_i \rangle) - c(p, \langle x_i, \hat{y} \rangle)| \alpha \tag{1.61}$$

$$= 0 \tag{1.62}$$

Since $m(x_i, \hat{y}_i, \hat{y}, w') > 0$ we have that $f_{w'}(x_i) \neq \hat{y}$ and the lemma follows. \blacksquare

1.7 Discussion

One of the goals of learning theory is to provide guidance in the construction of learning algorithms. This paper provides consistent generalization bounds for learning decoders (structured output learning). These new bounds improve previous bounds by achieving consistency for arbitrary distortion (loss) functions. These new generalization bounds seem to provide insight into the various notions of hinge loss that have been proposed for learning decoders and suggest that nonconvex optimization may achieve superior, or at least consistent, generalization.

Acknowledgments

I would like to thank Michael Collins, Ben Taskar and Peter Bartlett for useful discussions regarding this paper.

References

- [1] Y. Altun and T. Hofmann. Large margin methods for label sequence learning. In *8th European Conference on Speech Communication and Technology*, 2003.
- [2] Peter Bartlett, Michael Collins, Ben Taskar, and David McAllester. Exponentiated gradient algorithms for large-margin structured classification. In *NIPS 04*, 2004.
- [3] Michael Collins. Parameter estimation for statistical parsing models: Theory and practice of distribution-free methods. In *New Developments in Parsing Technology*. Kluwer Academic, 2004. Revised version of the paper that appeared at IWPT 2001.
- [4] L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In S. Dzeroski and N. Lavrac, editors, *Relational Data Mining*. Springer-Verlag, 2001.
- [5] Keiji Kanazawa, Daphne Koller, and Stuart Russell. Stochastic simulation algorithms for dynamic probabilistic networks. In *UAI95*, 1995.
- [6] J. D. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic modeling for segmenting and labeling sequence data. In *18th International Conference on Machine Learning ICML*, 2001.
- [7] John Langford and John Shawe-Taylor. PAC-Bayes and margins. In *Neural Information Processing Systems (NIPS)*, 2002.
- [8] Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99, 2004.
- [9] David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 5:5–21, 2003. A short version appeared as "PAC-Bayesian Model Averaging" in COLT99.
- [10] David McAllester. Simplified PAC-Bbayesian margin bounds. In *COLT03*, 2003.
- [11] Matthias Seeger. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002.
- [12] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *Neural Information Processing Systems 2003*, 2003.