Generalization Bounds for Learning with Linear and Quadratic Side Knowledge

Theja Tulabandhula and **Cynthia Rudin**
MIT, Cambridge MA 02139

Abstract

In this paper, we consider a supervised learning setting where side knowledge is provided about the labels of unlabeled examples. The side knowledge has the effect of reducing the hypothesis space, leading to tighter generalization bounds, and thus possibly better generalization. We consider two types of side knowledge, the first leading to linear constraints on the hypothesis space, and the second leading to quadratic constraints on the hypothesis space. We show how different types of domain knowledge can lead directly to these kinds of side knowledge. We prove bounds on complexity measures of the hypothesis space for quadratic side knowledge, and show that these bounds are tight in a specific sense.

1 Introduction

Surely, for many applications the amount of domain knowledge we could potentially use within our learning processes is vastly larger than the amount of domain knowledge we actually use. One reason for this is that domain knowledge may be nontrivial to incorporate into algorithms or analysis. A few types of domain knowledge that do permit analysis have been explored quite in depth in the past few years and used very successfully in a variety of learning tasks; this includes knowledge about the sparsity properties of linear models (ℓ_1-norm constraints, minimum description length) or smoothness properties (ℓ_2-norm constraints, maximum entropy). A reason that domain knowledge is not usually incorporated in theoretical analysis is that it can be very problem specific; it may be too specific to the domain to have an overarching theory of interest. For example, researchers in NLP (Natural Language Processing) have long figured out various exotic domain specific knowledge that one can use while performing a learning task (Chang, Ratinov, and Roth, 2008; Chang et al., 2008). The present work aims to provide theoretical guarantees for a large class of problems with a general type of domain knowledge that goes beyond sparsity and smoothness.

To define this large class of problems, we will keep the usual supervised learning assumption that the training examples are drawn i.i.d. Additionally in our setting, we have a different set of examples without labels, not necessarily chosen randomly. For this set of unlabeled examples, we have some prior knowledge about the relationships between their labels, which affects the space of hypotheses we are searching over within our learning algorithms. These assumptions can, for example, take into account our partial knowledge about how any learned model should predict on the unlabeled examples if they were encountered. We consider two types of side knowledge, namely constraints on the unlabeled examples leading to (i) linear constraints on a linear function class, and (ii) quadratic constraints on a linear function class. Our main contributions are:

- To show that linear and quadratic constraints on a linear hypothesis space can arise naturally in many circumstances, from constraints on a set of unlabeled examples. This is in Section 2. We connect these with relevant semi-supervised learning settings.
- To provide a bound on the complexity of the hypothesis space for the quadratic constraint case, which is tighter than previous results. This can be used directly in generalization bounds. Our bound is contained in Section 3. Bounds for the case of linear constraints are found in Section 3.2. The bounds in Section 3.2 are not original to this paper, but their application to general side knowledge with linear constraints is novel. The quadratic bounds of Section 3.3 are novel to this paper.
- To show that the upper bound on the hypothesis space we provided has a matching lower bound, also in Section 3.3.

Side knowledge can be particularly helpful in cases where data are scarce; these are precisely circumstances when data themselves cannot fully define the predictive model, and thus domain knowledge can make an impact in predictive accuracy. That said, for any type of side knowledge (sparsity, smoothness, and the side knowledge considered here), the examples and hypothesis space may not conform in reality to the side knowledge. (Similarly, the training data may not be truly random in practice.) However, if they do, we can claim lower sample complexities, and potentially improve our model selection efforts. Thus, we cannot claim that our side knowledge is always true knowledge, but we can claim that if it is true, we are able to gain some benefit in learning.

2 Linear and Quadratic Constraints

We are given training sample S of n examples $\{(x_i, y_i)\}_{i=1}^n$ with each observation x_i belong to a set $\mathcal{X}$ in $\mathbb{R}^p$. Let the label y_i belong to a set $\mathcal{Y}$ in $\mathbb{R}$. In addition, we are given a set of m unlabeled examples $\{\tilde{x}_i\}_{i=1}^m$. We are not given the true labels $\{\tilde{y}_i\}_{i=1}^m$ for these observations. Let $\mathcal{F}$ be the function class (set of hypotheses) of interest, from which we want to choose a function f to predict the label of future unseen observations. Let it be linear, parameterized by coefficient vector β and its description will change based on the constraints we place on β.

Consider the empirical risk minimization problem: $\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} l(f,\{x_i,y_i\})$. Regularization on f acts to enforce assumptions that the true model comes from a restricted class, so that $\mathcal{F}$ is now defined as

$$\{f|f:\mathcal{X}\mapsto\mathcal{Y}, f(x)=\beta^T x, R_l(f)\leq c_l \text{ for } l=1,...,L\},$$

where $()^T$ represents the transpose operation. Here we have appended L additional constraints for regularization to the description of the hypothesis set $\mathcal{F}$. Especially if the training set is small, side knowledge can be very powerful in reducing the size of $\mathcal{F}$. Particularly if constants $\{c_l\}_{l=1}^{L}$ are small, the size of $\mathcal{F}$ be reduced substantially.

2.1 Assumptions leading to linear constraints

We will provide three settings to demonstrate that linear constraints arise in a variety of natural settings: poset, must-link, and sparsity on $\{\tilde{y}_i\}_{i=1}^{m}$. In all three, we will include standard regularization of the form $\|\beta\|_q \leq c_1$ by default.

Poset: Partial order information about the labels $\{\tilde{y}_i\}_{i=1}^{m}$ can be captured via the following constraints: $f(\tilde{x}_i)\leq f(\tilde{x}_k)+c$ for any collection of pairs $(i,k)\in[1,...,m]\times[1,...,m]$. This gives us up to m^2 constraints of the form $\beta^T(\tilde{x}_i-\tilde{x}_k)\leq c$. $\mathcal{F}$ can be described as: $F := \{f|f(x)=\beta^T x, \|\beta\|_q \leq c_1, \beta^T(\tilde{x}_i-\tilde{x}_k)\leq c_{i,k}, \forall(i,k)\in E\}$, where E is the set of pairs of indices of unlabeled data that are constrained.

Must-link: Here we bound the absolute difference of labels between pairs of unlabeled examples: $|f(\tilde{x}_i)-f(\tilde{x}_j)|\leq c$. This captures knowledge about the nearness of the labels. This leads to two linear constraints: $-c\leq\beta^T(\tilde{x}_i-\tilde{x}_j)\leq c$. These constraints have been used extensively within the semi-supervised (Zhu, 2005) and constrained clustering settings (Lu and Leen, 2004; Basu et al., 2006) as must-link or 'in equivalence' constraints. For must-link constraints, $\mathcal{F}$ is defined as: $\mathcal{F} := \{f|f(x)=\beta^T x, \|\beta\|_q\leq c_1, -c_{i,j}\leq \beta^T(\tilde{x}_i-\tilde{x}_j)\leq c_{i,j}, \forall(i,j)\in E\}$, where E is again the set of pairs of indices of unlabeled data that are constrained.

Sparsity and its variants on a subset of $\{\tilde{y}_i\}_{i=1}^{m}$: Similar to sparsity assumptions on β, here we want that only a small set of labels is nonzero among a set of unlabeled examples. In particular, we want to bound the cardinality of the support of the vector $[\tilde{y}_1 \ldots \tilde{y}_{|\mathcal{I}|}]$ for some index set $\mathcal{I}\subset\{1,...,m\}$. Such a constraint is nonlinear. Nonetheless, a convex constraint of the form $\|[\tilde{y}_1\ldots\tilde{y}_{|\mathcal{I}|}]\|_1 \leq c$ ($2^{|\mathcal{I}|}$ linear constraints) can be used as a proxy to encourage sparsity. The function class is defined as: $\mathcal{F} := \{f|f(x)=\beta^T x, \|\beta\|_q\leq c_1, \|[\beta^T\tilde{x}_1\ldots\beta^T\tilde{x}_{|\mathcal{I}|}]\|_1\leq c\}$, A similar constraint can be obtained if we instead had a partial information about the dual norm $\|[\tilde{y}_1\ldots\tilde{y}_{|\mathcal{I}|}]\|_\infty$.

2.2 Assumptions leading to quadratic constraints

We will provide several settings to show that quadratic constraints arise naturally.

Must-link: A constraint of the form $(f(\tilde{x}_i)-f(\tilde{x}_j))^2\leq c$ can be written as $0\leq\beta^T A\beta\leq c$ with $A=(\tilde{x}_i-\tilde{x}_j)(\tilde{x}_i-\tilde{x}_j)^T$. Here A is rank-deficient as it is an outer product, which leads to an unbounded ellipse; however, its intersection with a full ellipsoid (for instance, an ℓ_2-norm ball) is not unbounded and indeed can be a restricted hypothesis set. Set $\mathcal{F}$ is defined by: $\mathcal{F}=\{\beta:\beta^T\beta\leq c_1, \beta^T(\tilde{x}_i-\tilde{x}_j)(\tilde{x}_i-\tilde{x}_j)^T\beta\leq c_{i,j}; (i,j)\in E\}$, where E is again the set of pairs of indices of unlabeled data that are constrained.

Constraining label values for a pair of examples: We can define the following relationship between the labels of two unlabeled examples using quadratic constraints: if one of them is large in magnitude, the other is necessarily small. This can be encoded using the inequality: $f(\tilde{x}_i)\cdot f(\tilde{x}_j)\leq c$. If $f(x)\in\mathcal{Y}\subset\mathbb{R}_+$, then $f(\tilde{x}_i)\cdot f(\tilde{x}_j)\leq c$ gives the following quadratic constraint on β with the associated rank 1 matrix being $A=\tilde{x}_i\tilde{x}_j^T$: $\beta^T A\beta\leq c$. This is not quite an ellipsoidal constraint yet because matrices associated with ellipsoids are symmetric positive semidefinite. Matrix A on the other hand is not symmetric. Nonetheless, the quadratic constraint remains intact when we replace matrix A with the symmetric matrix $\frac{1}{2}(A+A^T)$. If in addition, the symmetric matrix is also positive-definite (which can be verified easily), then this leads to an ellipsoidal constraint. The hypothesis space $\mathcal{F}$ becomes: $\mathcal{F}=\left\{\beta:\beta^T\beta\leq C_1, \beta^T\tilde{x}_i\tilde{x}_j^T\beta\leq c_{i,j}; (i,j)\in E\right\}$.

Energy of estimated labels: We can place an upper bound constraint on the sum of squares (the "energy") of the predictions, which is: $||X_U^T\beta||_2^2=\sum_i(\beta^T\tilde{x}_i)^2=\beta^T(\sum_i\tilde{x}_i\tilde{x}_i^T)\beta$ where X_U is a $p\times m$ dimensional matrix with $\tilde{x}_i$'s as its columns.[1] The set $\mathcal{F}$ is $\mathcal{F}=\left\{\beta:\beta^T\beta\leq c_1, ||X_U^T\beta||_2^2\leq c\right\}$. Extensions like having the norm act on only a subset of the estimates of $\{\tilde{y}\}_{i=1}^{m}$ follow accordingly.

Smoothness and other constraints on $\{\tilde{y}_i\}_{i=1}^{m}$: Consider the general ellipsoid constraint $\|\Gamma X_U^T\beta\|_2^2\leq c$ where we have added an additional transformation matrix Γ in front of $X_U^T\beta$. If Γ is set to the identity matrix, we get the energy constraint previously discussed. If Γ is a banded matrix with $\Gamma_{i,i}=1$ and $\Gamma_{i,i+1}=-1$ for all $i=1,...,m$ and remaining entries zero, then we are encoding the side knowledge that the variation in the labels of the unlabeled examples is smoothly varying: we are encouraging the unlabeled examples with neighboring indices to have similar predicted values. This matrix Γ is an instance of a difference operator in the numerical analysis literature. In this context, banded matrices like Γ model discrete derivatives. By including this type of constraint, problems with identifiability and ill-posedness of an optimal solution β are alleviated. That is, as with the Tikhonov regularization on β in least squares regression, constraints derived from matrices like Γ reduce the condition number. The set $\mathcal{F}$ is $\mathcal{F}=\left\{\beta:\beta^T\beta\leq c_1, \|\Gamma X_U^T\beta\|_2^2\leq c\right\}$.

Graph based methods: Some graph regularization methods such as manifold regularization (Belkin and Niyogi,

[1] Note that this notation is not the usual notation where observations $\tilde{x}_i$'s are stacked as rows.

2004) also encode information about the labels of the unlabeled data. They also lead to convex quadratic constraints on β. Here, along with the unlabeled examples $\{\tilde{x}_i\}_{i=1}^m$, our side knowledge consists of an m-node weighted graph $G = (V, E)$ with the Laplacian matrix $L_G = D - A$. Here, D is a $m \times m$-dimensional diagonal matrix with the diagonal entry for each node equal to the sum of weights of the edges connecting it. Further, A is the adjacency matrix containing the edge weights a_{ij}, where $a_{ij} = 0$ if $(i, j) \notin E$ and $a_{ij} = e^{-c\|\tilde{x}_i - \tilde{x}_j\|_q}$ if $(i, j) \in E$ (other choices for the weights are also possible). The quadratic function $(X_U^T \beta)^T L_G (X_U^T \beta)$ is then twice the sum over all edges, of the weighted squared difference between the two node labels corresponding to the edge: $2 \sum_{(i,j)\in E} a_{ij} \left(f(\tilde{x}_i) - f(\tilde{x}_j)\right)^2$. Intuitively, if we have the side knowledge that this quantity is small, it means that a node should have similar labels to its neighbors. For classification, this typically encourages the decision boundary to avoid dense regions of the graph. The set $\mathcal{F}$ is defined as: $\mathcal{F} = \{\beta : \beta^T \beta \leq c_1, \beta^T X_U^T L_G X_U^T \beta \leq c\}$.

3 Generalization Bounds

We know that every piece of additional information about the solution, including side information, can reduce the complexity of the hypothesis space and may thus promote generalization. In this section, we will make this precise. We will compute bounds on the complexity of the hypothesis space when the types of constraints seen in Section 2 are included.

3.1 Definition of Complexity Measures

We will look at two complexity measures: the covering number of a hypothesis set and the Rademacher complexity of a hypothesis set. Their definitions are as follows:

Definition 1. *Covering Number (Kolmogorov and Tikhomirov, 1959):* Let $A \subseteq \Omega$ be an arbitrary set and (Ω, ρ) a (pseudo-)metric space. Let $|\cdot|$ denote set size. For any $\epsilon > 0$, an ϵ**-cover** for A is a finite set $U \subseteq \Omega$ (not necessarily $\subseteq A$) s.t. $\forall \omega \in A, \exists u \in U$ with $d_\rho(\omega, u) \leq \epsilon$. The **covering number** of A is $N(\epsilon, A, \rho) := \inf_U |U|$ where U is an ϵ-cover for A.

Definition 2. *Rademacher Complexity (Bartlett and Mendelson, 2002):* Given a training sample $S = \{x_1, ..., x_n\}$, with each x_i drawn i.i.d. from $\mu_\mathcal{X}$, and hypothesis space $\mathcal{F}$, $\mathcal{F}_{|S}$ is the defined as the restriction of $\mathcal{F}$ with respect to S. The *empirical Rademacher complexity of* $\mathcal{F}_{|S}$ is

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right]$$

where $\{\sigma_i\}$ are Rademacher random variables ($\sigma_i = 1$ with probability $1/2$ and $\sigma_i = -1$ with probability $1/2$). The *Rademacher complexity of* $\mathcal{F}$ is its expectation:

$$\mathcal{R}(\mathcal{F}) = \mathbb{E}_{S \sim (\mu_\mathcal{X})^n} [\bar{\mathcal{R}}(\mathcal{F}_{|S})].$$

If instead we let $\sigma_i \sim \mathcal{N}(0, 1)$ in the definition, this is the Gaussian complexity of the function class. Generalization bounds often use both these quantities in their statements (Bartlett and Mendelson, 2002).

3.2 Complexity results with linear constraints

We will state two results: one for a single linear constraint and the other for multiple linear constraints. They were designed for a specific type of side knowledge, namely knowledge about the cost to solve a decision problem associated with the learning problem (Tulabandhula and Rudin, 2013b,a). The crux of our argument in Section 2.1 is that these bounds extend well beyond that.

Single linear constraint

Theorem 3.1. (Theorem 2 of Tulabandhula and Rudin, 2013b) Let $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_2 \leq X_b\}$ and $\mu_\mathcal{X}$ be the marginal probability measure on $\mathcal{X}$. Let $\mathcal{F} = \left\{f | f : \mathcal{X} \mapsto \mathcal{Y}, f(x) = \beta^T x, \|\beta\|_2 \leq B_b, a^T \beta \leq 1\right\}$. Further let $\mathcal{F}_{|S} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$. Then for all $\epsilon > 0$, for any sample S,

$$N(\epsilon / X_b, \mathcal{F}_{|S}, \|\cdot\|_2) \leq \alpha(p, a, \epsilon) c(p) \left(\frac{2 B_b X_b}{\epsilon} + 1\right)^p$$

where $c(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)}$. Also, defining $r = B_b + \frac{\epsilon}{2X_b}$ and $V_p(r) = c(p) r^p$, the function α above is:

$$\alpha(p, a, \epsilon) =$$
$$1 - \frac{1}{V_p(r)} \int_{\theta = \cos^{-1}\left(\frac{\|a\|_2^{-1} + \frac{\epsilon}{2X_b}}{r}\right)}^{0} V_{p-1}(r \sin\theta) d(r \cos\theta).$$

It is known (Kolmogorov and Tikhomirov, 1959) that $\mathcal{B} = \{\beta : \|\beta\|_2 \leq B_b\}$ has a bound on its covering number of the form $N(\epsilon, \mathcal{B}, \|\cdot\|_2) \leq c(p) \left(\frac{2B_b}{\epsilon} + 1\right)^p$ where $c(p) = \pi^{p/2} / \Gamma(\frac{p}{2} + 1)$. Since in Theorem 3.1 the same constant appears, multiplied by a factor at most one, the bound in Theorem 3.1 can be tighter. The function $\alpha(p, a, \epsilon)$ can be considered to be the normalized volume of the ball (which is 1) minus the portion that is the spherical cap cut off by the linear constraint. It comes directly from formulae for the volume of spherical caps. We are integrating over the volume of a $p - 1$ dimensional sphere of radius $r \sin\theta$ and the height term is $d(r \cos\theta)$.

This bound shows that the covering number bound can depend on a which is a direct function of the unlabeled examples $\{\tilde{x}_i\}_{i=1}^m$. As the norm $\|a\|_2$ increases, $\|a\|_2^{-1}$ decreases, thus $\alpha(p, a, \epsilon)$ decreases, and the whole bound decreases. This is a mechanism by which side information on the labels of the unlabeled examples influences the complexity measure of the hypothesis set, potentially improving generalization.

Multiple linear constraints and general norm constraints

Let us define the matrix $[x_1 \ \ldots \ x_n]$ as matrix X_L where $x_i \in \mathcal{X} = \{x : \|x\|_r \leq X_b\}$. Then, X_L^T can be written as $[h_1 \cdots h_p]$ with $h_j \in \mathbb{R}^n, j = 1, ..., p$. Define function class $\mathcal{F}$ as

$$\mathcal{F} = \Big\{ f | f(x) = \beta^T x, \beta \in \mathbb{R}^p, \|\beta\|_q \leq B_b,$$
$$\sum_{j=1}^p c_{j\nu} \beta_j + \delta_\nu \leq 1, \delta_\nu > 0, \nu = 1, ..., V \Big\},$$

where $1/r+1/q=1$ and $\{c_{j\nu}\}_{j,\nu}$, $\{\delta_\nu\}_\nu$ and B_b are known constants.

Let $\{\tilde{c}_{j\nu}\}_{j,\nu}$ be proportional to $\{c_{j\nu}\}_{j,\nu}$:

$$\tilde{c}_{j\nu} \quad := \quad \frac{c_{j\nu} n^{1/r} X_b B_b}{\|h_j\|_r} \quad \forall j = 1,...,p \text{ and } \nu = 1,...,V.$$

Let K be a positive number. Further, let the sets P^K parameterized by K and P_c^K parameterized by K and $\{\tilde{c}_{j\nu}\}_{j,\nu}$ be: $P^K := \left\{(k_1,...,k_p) \in \mathbb{Z}^p : \sum_{j=1}^p |k_j| \le K\right\}$, and $P_c^K := \left\{(k_1,...,k_p) \in P^K : \sum_{j=1}^p \tilde{c}_{j\nu} k_j \le K \; \forall \nu = 1,...,V\right\}$. Let $|P^K|$ and $|P_c^K|$ be the sizes of the sets P^K and P_c^K respectively. The subscript c in P_c^K denotes that this polyhedron is a constrained version of P^K. As the linear constraints given by the $c_{j\nu}$'s force the hypothesis space to be smaller, they force $|P_c^K|$ to be smaller. Define X_{sL} to be equal to a diagonal matrix whose j^{th} diagonal element is $\frac{n^{1/r} X_b B_b}{\|h_j\|_r}$ times X_L. Define $\lambda_{\min}(X_{sL} {X_{sL}}^T)$ to be the smallest eigenvalue of the matrix $X_{sL} {X_{sL}}^T$.

Theorem 3.2. (Theorem 6 of Tulabandhula and Rudin, 2013a)

$$N(\sqrt{n}\epsilon, \mathcal{F}_{|S}, \|\cdot\|_2) \le \begin{cases} \min\{|P^{K_0}|, |P_c^K|\} & \text{if } \epsilon < X_b B_b \\ 1 & \text{otherwise} \end{cases},$$

where $K_0 = \left\lceil \frac{X_b^2 B_b^2}{\epsilon^2} \right\rceil$ and K is the maximum of K_0 and

$$\left\lceil \frac{n X_b^2 B_b^2}{\lambda_{\min}(X_{sL} {X_{sL}}^T) \left[\min_{\nu=1,...,V} \frac{\delta_\nu}{\sum_{j=1}^p |\tilde{c}_{j\nu}|} \right]^2} \right\rceil.$$

The linear assumptions on the labels of the unlabeled examples $\{\tilde{x}_i\}_{i=1}^m$ determine the parameters $\{\tilde{c}_{j\nu}\}_{j,\nu}$ which in turn influence the complexity measure bound.

3.3 Complexity results for quadratic structure

Consider the set $\mathcal{F} = \{f : f = \beta^T x, \beta^T A_1 \beta \le 1, \beta^T A_2 \beta \le 1\}$. Assume that at least one of the matrices is positive definite and both are positive-semidefinite, symmetric. Let $\Xi_1 = \{\beta : \beta^T A_1 \beta \le 1\}$ and $\Xi_2 = \{\beta : \beta^T A_2 \beta \le 1\}$ be the corresponding ellipsoid sets.

We first find an ellipsoid $\Xi_{\text{int}\gamma}$ (with matrix $A_{\text{int}\gamma}$) circumscribing the intersection of the two ellipsoids Ξ_1 and Ξ_2 and then find a bound on the Rademacher complexity of a corresponding function class leading to our result for the quadratic constraint case. We will pick matrix $A_{\text{int}\gamma}$ to have a particularly desirable property, namely that it is *tight*. We will call a circumscribing ellipsoid *tight* when no other ellipsoidal boundary comes between its boundary and the intersection $(\Xi_1 \cap \Xi_2)$. If we thus choose this property as our criterion for picking the ellipsoid, then according to the following result, we can do so by a convex combination of the original ellipsoids:

Theorem 3.3. (Circumscribing ellipsoids Kahan, 1968) There is a family of circumscribing ellipsoids that contains every tight ellipsoid. Every ellipsoid $\Xi_{\text{int}\gamma}$ in this family has $\Xi_{\text{int}\gamma} \supseteq (\Xi_1 \cap \Xi_2)$ and is generated by matrix $A_{\text{int}\gamma} = \gamma A_1 + (1-\gamma) A_2$, $\gamma \in [0,1]$.

Using the above theorem, we can find a tight ellipsoid $\{\beta : \beta^T A_{\text{int}\gamma} \beta \le 1\}$ that contains the set $\{\beta : \beta^T A_1 \beta \le 1, \beta^T A_2 \beta \le 1\}$ easily. Note that the right hand sides of the quadratic constraints defining these ellipsoids can be equal to one without loss of generality.

Theorem 3.4. (Rademacher complexity of linear function class with two quadratic constraints) Let

$$\mathcal{F} = \{f : f(x) = \beta^T x : \beta^T \mathbb{I} \beta \le B_b^2, \beta^T A_2 \beta \le 1\}$$

with A_2 symmetric positive-semidefinite. Then,

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) \le \frac{1}{n} \sqrt{\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L)}, \qquad (1)$$

where $A_{\text{int}\gamma}$ is the matrix of a circumscribing ellipsoid $\{\beta : \beta^T A_{\text{int}\gamma} \beta \le 1\}$ of the set $\{\beta : \beta^T \mathbb{I} \beta \le B_b^2, \beta^T A_2 \beta \le 1\}$ and X_L is the matrix $[x_1 \; \ldots \; x_n]$ with examples x_i's as its columns.

Proof. Consider the set $\mathcal{F}_{|S} = \{(\beta^T x_1, ..., \beta^T x_n) \in \mathbb{R}^n : \beta^T \mathbb{I} \beta \le B_b^2, \beta^T A_2 \beta \le 1\} \subset \mathbb{R}^n$. Let $\sigma = [\sigma_1, ..., \sigma_n]^T$. Also, let $\alpha = A_{\text{int}\gamma}^{1/2} \beta$.

$$\begin{aligned}
\bar{\mathcal{R}}(\mathcal{F}_{|S}) &\overset{(a)}{\le} \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\{\beta : \beta^T A_{\text{int}\gamma} \beta \le 1\}} \sum_{i=1}^n \sigma_i \beta^T x_i \right] \\
&\overset{(b)}{=} \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\{\alpha : \alpha^T \alpha \le 1\}} \sum_{i=1}^n \sigma_i (A_{\text{int}\gamma}^{-1/2} \alpha)^T x_i \right] \\
&= \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{\{\alpha : \|\alpha\|_2 \le 1\}} \alpha^T (A_{\text{int}\gamma}^{-1/2})^T X_L \sigma \right] \\
&\overset{(c)}{=} \frac{1}{n} \mathbb{E}_\sigma \left[\|(A_{\text{int}\gamma}^{-1/2})^T X_L \sigma\|_2 \right] \\
&\overset{(d)}{\le} \frac{1}{n} \sqrt{\mathbb{E}_\sigma \left[\|(A_{\text{int}\gamma}^{-1/2})^T X_L \sigma\|_2^2 \right]} \\
&= \frac{1}{n} \sqrt{\mathbb{E}_\sigma \left[\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L \sigma \sigma^T) \right]} \\
&\overset{(e)}{=} \frac{1}{n} \sqrt{\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L)}
\end{aligned}$$

where (a) follows because we are taking the supremum over the circumscribing ellipsoid; (b) follows because $A_{\text{int}\gamma}$ is positive definite, hence invertible; (c) is by Cauchy-Schwarz (equality case); (d) uses Jensen's inequality and (e) uses the linearity of trace and expectation to commute them along with the fact that $\mathbb{E}[\sigma \sigma^T] = I$. □

If $A_{\text{int}\gamma}$ is diagonal (or axis-aligned), then we can write the empirical complexity $\bar{\mathcal{R}}(\mathcal{F}_{|S})$ in terms of the eigenvalues $\{\lambda_i\}_{i=1}^p$ as $\bar{\mathcal{R}}(\mathcal{F}_{|S}) \le \frac{1}{n} \sqrt{\sum_{j=1}^n \sum_{i=1}^p \frac{x_{ji}^2}{\lambda_i}}$ and this can be bounded by $\frac{X_b B_b}{\sqrt{n}}$ (Kakade, Sridharan, and Tewari, 2008) when $A_2 = \mathbf{0}$. In that case, all of the λ_i are $\frac{1}{B_b^2}$.

Since we can choose any circumscribing matrix $A_{\text{int}\gamma}$ in this theorem, we can perform the following optimization to get a circumscribing ellipsoid that minimizes the bound:

$$\min_{\gamma\in[0,1]} \text{trace}(X_L^T(\gamma A_1+(1-\gamma)A_2)^{-1}X_L).$$

This optimization problem is a univariate non-linear program. Again, as discussed extensively in Section 2, the matrix A_2 can be a function of the unlabeled data, encoding a variety of side knowledge.

We will now show that the dependence of the complexity on the sum of the inverse eigenvalues of $A_{\text{int}\gamma}$ is near optimal. In order to do so, we will make use of the Gaussian complexity measure instead of the Rademacher complexity which is related to the former as follows.

Lemma 3.5. (Lemma 4 of Bartlett and Mendelson, 2002) There are absolute constants C and D such that for every $\mathcal{F}_{|S}$ with $|S|=n$,

$$D\bar{\mathcal{R}}(\mathcal{F}_{|S})\le\bar{\mathcal{G}}(\mathcal{F}_{|S})\le C\log(n)\bar{\mathcal{R}}(\mathcal{F}_{|S}).$$

Since $A_{\text{int}\gamma}$ is a real symmetric matrix, we can decompose $A_{\text{int}\gamma}$ into a product P^TDP where D is a diagonal matrix with the eigenvalues of $A_{\text{int}\gamma}$ as its entries and P is an orthogonal matrix (i.e., $P^TP=I$). Our result of the form of the bound of Theorem 3.4 is as follows.

Theorem 3.6.

$$\bar{\mathcal{R}}(\mathcal{F}_{|S})\ge\frac{\kappa}{n\log n}\sqrt{\text{trace}(X_L^TA_{\text{int}\gamma}^{-1}X_L)}$$

where

$$\kappa=\frac{2}{C\sqrt{1+\frac{2\pi pnX_b^2}{(\min_{j=1,\dots,p}\|(PX_L)_j\|_2)^2}}},$$

C is the constant in Lemma 3.5, P is the orthogonal matrix from the decomposition of $A_{\text{int}\gamma}$, p, X_b are problem constants and n is the number of training examples.

The proof for the lower bound is similar to what one would do for estimating the complexity of a ellipsoid itself (without regard to a corresponding linear function class). See also (Wainwright, 2011) for handling single ellipsoids.

Proof of Theorem 3.6:

Let us define a new variable: $\alpha:=P\beta$, which is a linear transformation of linear model parameter β. Then, the scaled Gaussian complexity of our function class obeys the following,

$$n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S})=\mathbb{E}_\sigma\left[\sup_{\alpha^TD\alpha\le1}\sum_{i=1}^n\sigma_i\alpha^TPx_i\right],$$

where $\{\sigma_i\}_{i=1}^n$ are i.i.d. standard normal random variables. We now define a new vector ω to be a transformed version of the random vector $\sum_{i=1}^n\sigma_ix_i$. That is, let $\omega(\sigma):=P\sum_{i=1}^n\sigma_ix_i$. We will drop the dependence of ω on σ from the notation when it is clear from the context. The expression now becomes

$$n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S})=\mathbb{E}_\sigma\left[\sup_{\alpha^TD\alpha\le1}\alpha^T\omega\right]. \tag{2}$$

We will need the following lemma which describes concentration for Lipschitz functions of gaussian random variables.

Lemma 3.7. (Concentration (Tsirelson, Ibragimov, and Sudakov, 1976)) If σ is a vector with i.i.d. standard normal entries and G is any function with Lipschitz constant $\mathcal{L}$ (with respect to the Euclidean norm), then

$$\mathbb{P}[|(G(\sigma)-\mathbb{E}[G(\sigma)]|\ge t]\le2e^{-\frac{t^2}{2\mathcal{L}^2}}.$$

We will now state three claims.

Note 1: The function $F(\omega):=\sup_{\alpha^TD\alpha\le1}\alpha^T\omega(\sigma)$ is Lipschitz in σ with a Lipschitz constant $\mathcal{L}$ bounded by $X_b\sqrt{\frac{p\cdot n}{\lambda_{min}(D)}}$.

Note 2: The mean of $F(\omega)$ is $\mathbb{E}_\sigma[F(\omega)]=n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S})$.

Note 3: For a random variable Y^2, $\mathbb{E}(Y^2)=\int_0^{+\infty}P(Y^2\ge s)ds$.

By *Notes 1* and *2*, and using Lemma 3.7 with $G(\sigma)=F(\omega)$, we have

$$\mathbb{P}[|(F(\omega)-\mathbb{E}_\sigma[F(\omega)]|\ge t]\le2e^{-\frac{t^2}{2\mathcal{L}^2}},$$

where $\mathcal{L}=X_b\sqrt{\frac{p\cdot n}{\lambda_{min}(D)}}$. Now we can bound the variance of $F(\omega)$ as follows. Let $Y=|(F(\omega)-\mathbb{E}_\sigma[F(\omega)]|$. Then from the above tail bound, $P(Y^2\ge s)\le2e^{-\frac{s}{2\mathcal{L}^2}}$ is also true. The variance of $F(\omega)$ which is the same as the expectation of Y^2 can thus be obtained as:

$$\begin{aligned}\text{Var}(F(\omega))=\mathbb{E}_\sigma(Y^2)&\overset{(*)}{=}\int_0^{+\infty}P(Y^2\ge s)ds\\&\le2\int_0^{+\infty}e^{-\frac{s}{2\mathcal{L}^2}}ds=4X_b^2\frac{p\cdot n}{\lambda_{min}(D)},\end{aligned}\tag{3}$$

where we substituted $X_b\sqrt{\frac{p\cdot n}{\lambda_{min}(D)}}$ for $\mathcal{L}$. For Equation (*) we used *Note 3*.

This upper bound on the variance of $F(\omega)$ is used to lower bound Rademacher complexity as follows. First we will lower bound the related Gaussian complexity by constructing a feasible candidate α' to substitute for the sup operation in Equation (2). Then we will use the variance upper bound on $F(\omega)$.

Lower bounding Gaussian complexity: Let $j^*\in\{1,\dots,p\}$ be the index at which the diagonal element $D(j^*,j^*)=\lambda_{min}(D)$. For each realization of σ (or equivalently ω) let $\alpha'=\left[0\dots\frac{|\omega_{j^*}|}{\omega_{j^*}\sqrt{\lambda_{min}(D)}}\dots0\right]$ with the non-zero entry at coordinate j^*. Clearly α' is a feasible vector in the ellipsoidal constraint $\{\alpha:\alpha^TD\alpha\le1\}$ seen in the complexity expression, Equation (2). Substituting it and using the definition of $F(\omega)$, we get a lower bound on the complexity:

$$\begin{aligned}n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S})=&\mathbb{E}_\sigma[F(\omega)]=\mathbb{E}_\sigma\left[\sup_{\alpha^TD\alpha\le1}\alpha^T\omega\right]\\&\overset{(a)}{\ge}\mathbb{E}_\sigma[(\alpha')^T\omega]\overset{(b)}{\ge}\frac{1}{\sqrt{\lambda_{min}(D)}}\mathbb{E}_\sigma[|\omega_{j^*}|].\end{aligned}$$

Step (a) comes from the fact that α' is feasible in $\{\alpha : \alpha^T D\alpha \leq 1\}$ but not necessarily the maximum, and step (b) comes from the definition of α'.

Note that compared to the upper bound on the related Rademacher complexity obtained in Theorem 3.4, the dependence on $A_{\text{int}\gamma}$ is weak (only via $\lambda_{min}(D)$). We will use the variance of $F(\omega)$ to obtain a lower bound very similar to the upper bound in Equation (1). Rearranging the terms in the previous inequality, we get:

$$\frac{(\mathbb{E}_\sigma[F(\omega)])^2}{(\mathbb{E}_\sigma|\omega_{j^*}|)^2} \geq \frac{1}{\lambda_{min}(D)}. \tag{4}$$

By rewriting the variance in terms of the second and first moments, using expression (3) and then using (4) we get

$$\begin{aligned}\text{Var}(F(\omega)) =& \mathbb{E}_\sigma[F^2(\omega)] - (\mathbb{E}_\sigma[F(\omega)])^2 \\ \leq& 4X_b^2 \frac{p\cdot n}{\lambda_{min}(D)} \leq 4pnX_b^2 \frac{(\mathbb{E}_\sigma[F(\omega)])^2}{(\mathbb{E}_\sigma|\omega_{j^*}|)^2}.\end{aligned}$$

Using expression (2) again, and then rearranging the terms in the previous expression, we obtain another lower bound on the scaled Gaussian complexity which is:

$$\begin{aligned}\left(n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S})\right)^2 =& (\mathbb{E}_\sigma[F(\omega)])^2 \geq \frac{\mathbb{E}_\sigma[(F(\omega))^2]}{1+\frac{4pnX_b^2}{(\mathbb{E}_\sigma|\omega_{j^*}|)^2}} \\ =& \frac{\mathbb{E}_\sigma[(\sup_{\alpha^T D\alpha\leq 1}\omega^T\alpha)^2]}{1+\frac{4pnX_b^2}{(\mathbb{E}_\sigma|\omega_{j^*}|)^2}}.\end{aligned} \tag{5}$$

We can now try to bound two easier quantities $\mathbb{E}_\sigma[(\sup_{\alpha^T D\alpha\leq 1}\omega^T\alpha)^2]$ and $\mathbb{E}_\sigma|\omega_{j^*}|$ to get an expression for scaled Gaussian complexity and consequently for Rademacher complexity.

Let us start first with $\mathbb{E}|\omega_{j^*}|$. By definition ω equals $PX_L\sigma$. Thus, the j^*th coordinate of ω will be $\sum_i \sigma_i(Px_i)_{j^*}$ where $(\cdot)_{j^*}$ represents the j^*th coordinate of the vector. Since the σ_i are independent standard normal, their weighted sum ω is also standard normal with variance $\sum_i (Px_i)^2_{j^*}$. Since for any normal random variable z with mean zero and variance d it is true that $\mathbb{E}[|z|] = \sqrt{\frac{2d}{\pi}}$, we have

$$\begin{aligned}\mathbb{E}_\sigma[|w_{j^*}|] =& \sqrt{\frac{2}{\pi}}\left(\sum_i (Px_i)^2_{j^*}\right)^{\frac{1}{2}} \\ \geq& \sqrt{\frac{2}{\pi}} \min_{j=1,\ldots,p} \|(PX_L)_j\|_2\end{aligned} \tag{6}$$

where $(PX_L)_j$ represents the j^{th} row of the matrix PX_L. For the second moment term of (5) that we need to bound, $\mathbb{E}_\sigma[(\sup_{\alpha^T D\alpha\leq 1}\omega^T\alpha)^2]$, we can see that

$$\begin{aligned}\sup_{\alpha^T D\alpha\leq 1}\omega^T\alpha =& \sup_{\tilde{\alpha}^T\tilde{\alpha}\leq 1}(PX_L\sigma)^T D^{-1/2}\tilde{\alpha} \\ =& \|D^{-1/2}PX_L\sigma\|_2.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}_\sigma\Big[\Big(\sup_{\alpha^T D\alpha\leq 1}\omega^T\alpha\Big)^2\Big] =& \mathbb{E}_\sigma[\|D^{-1/2}PX_L\sigma\|_2^2] \\ =& \mathbb{E}_\sigma[\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L\sigma\sigma^T)] \\ =& \text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L).\end{aligned} \tag{7}$$

Substituting the two bounds we just derived, (6) and (7), into (5) gives us a lower bound on the scaled Gaussian complexity:

$$\left(n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S})\right)^2 \geq \frac{\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L)}{1+\frac{4pnX_b^2}{(\sqrt{\frac{2}{\pi}}\min_{j=1,\ldots,p}\|(PX_L)_j\|_2)^2}}$$

$$n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S}) \geq \sqrt{\frac{\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L)}{1+\frac{4pnX_b^2}{(\sqrt{\frac{2}{\pi}}\min_{j=1,\ldots,p}\|(PX_L)_j\|_2)^2}}}.$$

Using Lemma 3.5 gives:

$$nC\log(n)\bar{\mathcal{R}}(\mathcal{F}_{|S}) \geq \sqrt{\frac{\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L)}{1+\frac{4pnX_b^2}{(\sqrt{\frac{2}{\pi}}\min_{j=1,\ldots,p}\|(PX_L)_j\|_2)^2}}}$$

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) \geq \frac{\kappa}{n\log n}\sqrt{\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L)}$$

where

$$\kappa = \frac{2}{C\sqrt{1+\frac{2\pi pnX_b^2}{(\min_{j=1,\ldots,p}\|(PX_L)_j\|_2)^2}}}.$$

In summary, we have proved matching lower bounds for the Rademacher complexity of quadratically constrained linear function classes.

4 Related Work

It is well-known that having additional unlabeled examples can aid in learning (Fung, Mangasarian, and Shavlik, 2002; Shental et al., 2004; Nguyen and Caruana, 2008), and this has been the subject of research in semi-supervised learning (Zhu, 2005). The present work is fundamentally different than semi-supervised learning, because semi-supervised learning exploits the distributional properties of the set of unlabeled examples. In this work, we do not necessarily have enough unlabeled examples to study these distributional properties, but these unlabeled examples do provide us information about the hypothesis space. Distributional properties used in semi-supervised learning include cluster assumptions (Singh, Nowak, and Zhu, 2008; Rigollet, 2007) and manifold assumptions (Belkin and Niyogi, 2004). In our work, the information we get from the unlabeled examples allows us to restrict the hypothesis space, which lets us be in the framework of empirical risk minimization and give theoretical generalization bounds via complexity measures of the restricted hypothesis spaces (Bartlett and Mendelson, 2002; Vapnik, 1998). While the focus of many works (e.g., Zhang, 2002; Maurer, 2006) is on complexity measures for ball-like function classes, our hypothesis spaces are more complicated, and arise here from constraints on the data.

In a different framework, that of Valiant's PAC learning, there are concentration statements about the risks in the presence of unlabeled examples (Balcan and Blum, 2005; Kääriäinen, 2005), though in these results, the unlabeled points are used in a very different way than in our work. While their results focus on exploiting unlabeled data to estimate distribution dependent quantities, our technology focuses on exploiting unlabeled data to restrict the hypothesis space directly.

5 Conclusion

In this paper, we have outlined how various additional information one might have about a learning problem can effectively help in generalization.We focused our attention on two types of additional information, one leading to linear constraints and the other leading to quadratic constraints, giving motivating examples and deriving complexity measure bounds. This work goes beyond the traditional paradigm of ball-like hypothesis spaces to study more exotic, yet realistic, hypothesis spaces, and is a starting point for more work on other interesting hypothesis spaces.

Acknowledgements

We wish to kindly thank Adel Basli for helpful discussions and the anonymous reviewers for their suggestions. T. Tulabandhula was supported during this work by an International Fulbright Science and Technology Award. C. Rudin was supported by NSF-CAREER IIS-1053407.

References

Balcan, M., and Blum, A. 2005. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of Conference on Learning Theory*. Springer. 69–77.

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

Basu, S.; Bilenko, M.; Banerjee, A.; and Mooney, R. J. 2006. Probabilistic semi-supervised clustering with constraints. In *Semi-supervised learning*. Cambridge, MA. MIT Press. 71–98.

Belkin, M., and Niyogi, P. 2004. Semi-supervised learning on riemannian manifolds. *Machine Learning* 56(1):209–239.

Chang, M.-W.; Ratinov, L.-A.; Rizzolo, N.; and Roth, D. 2008. Learning and inference with constraints. In *AAAI*, 1513–1518.

Chang, M.; Ratinov, L.; and Roth, D. 2008. Constraints as prior knowledge. In *ICML Workshop on Prior Knowledge for Text and Language Processing*, 32–39.

Fung, G. M.; Mangasarian, O. L.; and Shavlik, J. W. 2002. Knowledge-based support vector machine classifiers. In *Proceedings of Neural Information Processing Systems*, 521–528.

Kääriäinen, M. 2005. Generalization error bounds using unlabeled data. In *Proceedings of Conference on Learning Theory*. Springer. 127–142.

Kahan, W. 1968. Circumscribing an ellipsoid about the intersection of two ellipsoids. *Canadian Mathematical Bulletin* 11(3):437–441.

Kakade, S.; Sridharan, K.; and Tewari, A. 2008. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Proceedings of Neural Information Processing Systems* 22.

Kolmogorov, A. N., and Tikhomirov, V. M. 1959. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk* 14(2):3–86.

Lu, Z., and Leen, T. K. 2004. Semi-supervised learning with penalized probabilistic clustering. In *Proceedings of Neural Information Processing Systems*, 849–856.

Maurer, A. 2006. The Rademacher complexity of linear transformation classes. In *Proceedings of Conference on Learning Theory*. Springer. 65–78.

Nguyen, N., and Caruana, R. 2008. Improving classification with pairwise constraints: a margin-based approach. In *Machine Learning and Knowledge Discovery in Databases*. Springer. 113–124.

Rigollet, P. 2007. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research* 8:1369–1392.

Shental, N.; Bar-Hillel, A.; Hertz, T.; and Weinshall, D. 2004. Computing Gaussian mixture models with EM using equivalence constraints. In *Proceedings of Neural Information Processing Systems*, volume 16, 465–472.

Singh, A.; Nowak, R.; and Zhu, X. 2008. Unlabeled data: Now it helps, now it doesn't. In *Proceedings of Neural Information Processing Systems*, 1513–1520.

Tsirelson, B. S.; Ibragimov, I. A.; and Sudakov, V. N. 1976. Norms of gaussian sample functions. In *Proceedings of the Third Japan–U.S.S.R. Symposium on Probability Theory. Lecture Notes in Math.*, volume 550, 20–41. Springer.

Tulabandhula, T., and Rudin, C. 2013a. Machine learning with operational costs. *Journal of Machine Learning Research* 14:1989–2028.

Tulabandhula, T., and Rudin, C. 2013b. On combining machine learning with decision making. work in progress.

Vapnik, V. N. 1998. *Statistical learning theory*, volume 2. Wiley New York.

Wainwright, M. 2011. *Metric entropy and its uses (Chapter 3).* Unpublished draft.

Zhang, T. 2002. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research* 2:527–550.

Zhu, X. 2005. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.