Mach Learn (2015) 100:183–216
DOI 10.1007/s10994-014-5478-4

Generalization bounds for learning with linear, polygonal, quadratic and conic side knowledge

Theja Tulabandhula · Cynthia Rudin

Received: 15 December 2013 / Accepted: 24 November 2014 / Published online: 17 December 2014
© The Author(s) 2014

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 several types of side knowledge, the first leading to linear and polygonal constraints on the hypothesis space, the second leading to quadratic constraints, and the last leading to conic constraints. 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 and conic side knowledge, and show that these bounds are tight in a specific sense for the quadratic case.

Keywords Statistical learning theory · Generalization bounds · Rademacher complexity · Covering numbers, constrained linear function classes · Side knowledge

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

---

Editors: Toon Calders, Rosa Meo, Floriana Esposito, and Eyke Hullermeier.

T. Tulabandhula (✉)
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: theja@mit.edu

C. Rudin
MIT Sloan School of Management, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA
e-mail: rudin@mit.edu

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 et al. 2008a, b). 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. We motivate this knowledge as being obtained from domain experts. 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 many types of side knowledge, namely constraints on the unlabeled examples leading to (1) linear constraints on a linear function class, (2) quadratic constraints on a linear function class, and (3) conic constraints on a linear function class. Our main contributions are:

- To show that linear, polygonal, quadratic and conic constraints on a linear hypothesis space can arise naturally in many circumstances, from constraints on a set of unlabeled examples. This is in Sect. 2. We connect these with relevant semi-supervised learning settings.
- To provide upper bounds on covering number and empirical Rademacher complexity for linearly constrained linear function classes. Bounds for the case of linear and polygonal constraints are found in Sects. 3.3 and 3.4 respectively. Two of the three bounds in these sections are not original to this paper, but their application to general side knowledge with linear constraints is novel.
- To provide two upper bounds on the complexity of the hypothesis space for the quadratic constraint case This can be used directly in generalization bounds. The use of a certain family of circumscribing ellipsoids and the quadratic bounds of Sect. 3.5 are novel to this paper.
- To show that one of the upper bounds on the quadratically constrained hypothesis space we provided has a matching lower bound, also in Sect. 3.5. This is novel to this paper.
- To provide a bound on the complexity of the hypothesis space for the conic constraint case. These bounds are in Sect. 3.7 and are novel to this paper.
- We develop a novel proof technique for upper bounding linear, quadratic and conic constraint cases based on convex duality.

Figure 1 illustrates the various types of side knowledge.

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

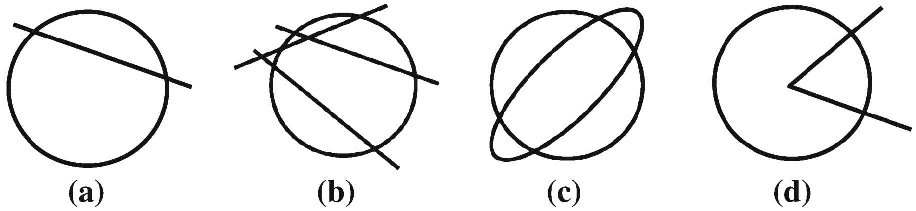

Fig. 1 This figure illustrates constraints on our hypothesis space. These constraints arise from side knowledge available about a set of unlabeled examples. The ℓ_2 balls in (**a**), (**b**), (**c**) and (**d**) represent coefficients of linear functions in two dimensions. **a** and **b** represent intersection of a ball and one or several half spaces. Theorems 1, 2 and Proposition 1 analyze these situations. **c** shows the intersection of a ball and an ellipsoid. Theorems 4, 5 and 6 correspond to this setting. (d) shows the intersection of a ball with a second order cone. Theorem 7 corresponds to this setting

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.

1.1 Motivating examples

Fung et al. (2002) added multiple linear constraints (polygonal constraints) to a specific ERM algorithm, the linear SVM, as a way to incorporate prior knowledge. They investigated the effect of using this type of prior knowledge for classification on a DNA promoter recognition dataset (Towell et al. 1990). In this classification task, the linear constraints result from precomputed rules that are separate from the training data (this is similar to our polygonal setting where constraints are generated from knowledge about the unlabeled examples). The “leave-one-out” error from the 1-norm SVM with the additional constraints was less than that of the plain 1-norm SVM and other training-data-based classifiers such as decision trees and neural networks. This and other types of knowledge incorporation in SVMs are reviewed by Lauer and Bloch (2008) and also Le et al. (2006).

James et al. (2014) motivated the use of linear constraints with LASSO, which is also an ERM procedure. In their experiment, they estimated a demand probability function using an on-line auto lending dataset. They ensured monotonicity of the demand function by applying a set of linear constraints (similar to the poset constraints in 2.1) and compared the output to two other methods: logistic regression and the unconstrained LASSO, both of which output non-monotonic demand probability curves.

Nguyen and Caruana (2008a) considered additional unlabeled examples whose labels are partially known. In particular, they worked on a type of multi-class classification task where they know that the label of each unlabeled example belongs to a known subset of the set of all class labels. This knowledge about the unlabeled examples translates into multiple linear constraints (polygonal constraints). They provided experimental results on five datasets showing improvements over multi-class SVMs.

Gómez-Chova et al. (2008) implemented a technique (known as LapSVMs) that uses Laplacian regularization augmented with standard SVMs for two image classification tasks related to urban monitoring and cloud screening (which are both remote sensing tasks). Laplacian regularization means that the regularization term is a quadratic function of the model, derived from a set of unlabeled examples, like our quadratic setting (see Sect. 2.2). In both tasks, the Laplacian-regularized linear SVMs outperformed the standard SVMs in terms of overall accuracy (these improvements are of the order of 2–3 % in both cases).

Shivaswamy et al. (2006) formulated robust classification and regression problems as described in Sect. 2.3 leading to conic constraints on the model class. For classification, they used the OCR, Heart, Ionosphere and Sonar datasets from the UCI repository to illustrate the effect of missing values and how robust SVM classification (which introduces second order conic constraints) provides better classification accuracy than the standard SVM classifier after imputation. For regression, they showed improvements in prediction accuracy of a robust version of SVR (again introducing conic constraints on the hypothesis space) as compared to a standard SVR trained after imputation on the Boston housing dataset (also from the UCI repository).

Finally, "Appendix" also provides experimental results showing the advantage of using side knowledge in a ridge regression problem.

2 Linear, polygonal, quadratic and conic 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: $\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i)$. Here the loss function is a Lipschitz continuous function such as the squared, exponential or hinge loss among others. This supervised learning setup encompasses both supervised classification ($\mathcal{Y}$ is a discrete set) and regression ($\mathcal{Y}$ is equal to $\mathbb{R}$). 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 and polygonal 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.

2.1.1 Poset

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

2.1.2 *Must-link*

Here we bound the absolute difference of labels between pairs of unlabeled examples: $|f(\tilde{x}_i) - f(\tilde{x}_j)| \leq c_{i,j}$. This captures knowledge about the nearness of the labels. This leads to two linear constraints: $-c_{i,j} \leq \beta^T(\tilde{x}_i - \tilde{x}_j) \leq c_{i,j}$. 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.

2.1.3 *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 $\|[\tilde{y}_1 \ldots \tilde{y}_{|\mathcal{I}|}]\|_1 \leq c_{\mathcal{I}}$ ($2^{|\mathcal{I}|}$ linear constraints) can be used as a proxy to encourage sparsity. The function class is defined as: $\mathcal{F} := \{f|f(x) = \beta^T x, \|\beta\|_q \leq c_1, \|[\beta^T\tilde{x}_1 \ldots \beta^T\tilde{x}_{|\mathcal{I}|}]\|_1 \leq c_{\mathcal{I}}\}$. A similar constraint can be obtained if we instead had partial information with respect to the dual norm: $\|[\tilde{y}_1 \ldots \tilde{y}_{|\mathcal{I}|}]\|_\infty \leq c_{\mathcal{I}}$.

2.2 Assumptions leading to quadratic constraints

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

2.2.1 *Must-link*

A constraint of the form $(f(\tilde{x}_i) - f(\tilde{x}_j))^2 \leq c_{i,j}$ can be written as $0 \leq \beta^T A\beta \leq c_{i,j}$ 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.

2.2.2 *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: $f(\tilde{x}_i) \cdot f(\tilde{x}_j) \leq c_{i,j}$. If $f(x) \in \mathcal{Y} \subset \mathbb{R}_+$, then $f(\tilde{x}_i) \cdot f(\tilde{x}_j) \leq c_{i,j}$ gives the following quadratic constraint on β with the associated rank 1 matrix being $A = \tilde{x}_i\tilde{x}_j^T$: $\beta^T A\beta \leq c_{i,j}$. 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: $\mathcal{F} = \left\{\beta : \beta^T\beta \leq c_1, \beta^T\tilde{x}_i\tilde{x}_j^T\beta \leq c_{i,j}; (i,j) \in E\right\}$.

2.2.3 *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 the use of Mahalanobis distance or having the norm act on only a subset of the estimates of $\{\tilde{y}\}_{i=1}^m$ follow accordingly.

2.2.4 *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 defined as: $\mathcal{F} = \left\{\beta : \beta^T\beta \leq c_1, \|\Gamma X_U^T\beta\|_2^2 \leq c\right\}$.

2.2.5 *Graph based methods*

Some graph regularization methods such as manifold regularization (Belkin and Niyogi 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: $\mathcal{F} = \{\beta : \beta^T\beta \leq c_1, \beta^T X_U^T L_G X_U^T\beta \leq c\}$.

2.3 Assumptions leading to conic constraints

We provide two scenarios that naturally lead to conic constraints on the model class: robustness against uncertainty and stochastic constraints.

2.3.1 *Robustness against uncertainty in linear constraints*

Consider any of the linear constraints considered in Sect. 2.1. All of these can be generically represented as: $\{a_k^T\beta \leq 1 \;\; \forall k = 1, .., K\}$ where for each k, a_k is a function of the unlabeled sample $\{\tilde{x}_j\}_{j=1}^m$ (for instance, $a_k = \tilde{x}_i - \tilde{x}_k$ for Poset constraints). Further assume that each a_k is only known to belong to an ellipsoid $\Xi_k = \{\overline{a}_k + A_k u : u^T u \leq 1\}$ with both parameters $\overline{a}_k$ and A_k known. This can happen due to measurement limitations, noise and other factors. We want to guarantee that, irrespective of the true value of $a_k \in \Xi_k$, we still have $a_k^T \beta \leq 1$.

---

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

Borrowing a trick used in the robust linear programming literature, we can encode (Lanckriet et al. 2003) the above requirement succinctly as:

$$\overline{a}_k^T \beta + \|A_k^T \beta\|_2 \leq 1, \forall k = 1, \ldots, K$$

which is a set of second-order cone constraints. The feasible set becomes smaller when the linear constraints $\{a_k^T \beta \leq 1 \ \forall k = 1, \ldots, K\}$ are replaced with the conic constraints above.

2.3.2 Stochastic Programming

Consider a probabilistic constraint of the form $\mathbb{P}_{a_k}(a_k^T \beta \leq 1) \geq \eta_k$, where a_k is now considered a random vector. The motivation for a_k is the same as before (see Sect. 2.1). If we know that a_k is normally distributed (for instance, due to additive noise) with mean $\overline{a}_k$ and covariance matrix B_k, then the probabilistic constraint is the same as: $\overline{a}_k^T \beta + \Phi^{-1}(1 - p)\|B_k^{1/2}\beta\|_2 \leq 1$, where $\Phi^{-1}()$ is the inverse error function. To see this, let $u_k = a_k^T \beta$ be a scalar random variable with mean $\overline{u_k}$ and variance σ_k^2 (this is equal to $\beta^T B_k \beta$). Then, our original constraint can be written as $\mathbb{P}\left(\frac{u_k - \bar{u}_k}{\sigma_k} \leq \frac{1 - \bar{u}_k}{\sigma_k}\right) \geq \eta_k$. Since $\frac{u_k - \bar{u}_k}{\sigma_k} \sim \mathcal{N}(0, 1)$, we can rewrite our constraint as: $\Phi\left(\frac{1 - \bar{u}_k}{\sigma_k}\right) \geq \eta_k$ where $\Phi(z)$ is the cumulative distribution function for the standard normal. Further $\Phi\left(\frac{1 - \bar{u}_k}{\sigma_k}\right) \geq \eta_k$ implies $\frac{1 - \bar{u}_k}{\sigma_k} \geq \Phi^{-1}(\eta_k)$. Rearranging terms, we get $\bar{u}_k + \Phi^{-1}(\eta_k)\sigma_k \leq 1$. Finally, substituting the values for $\bar{u}_k$ and σ_k gives us the following constraint:

$$\overline{a}_k^T \beta + \Phi^{-1}(\eta_k)\|B_k^{1/2}\beta\|_2 \leq 1,$$

which is a second order conic constraint (Lobo et al. 1998).

Remark 1 A question of practical interest would be about ways to impose constraints seen in Sects. 2.1, 2.2 and 2.3 in a computationally efficient manner. Fortunately, for all the cases we have considered thus far, the side knowledge can be encoded as a set of convex constraints leading to efficient algorithms (if the original empirical risk minimization problem is convex). Further, note that unlike must-link and similarity side knowledge that lead to convex constraints, cannot-link and dissimilarity knowledge is relatively harder to impose and is typically non-convex.

3 Generalization bounds

In each of the scenarios considered in Sect. 2, essentially we are given m unlabeled examples $\tilde{x}$ whose subsets satisfy various properties or side knowledge (for instance, linear ordering, quadratic neighborhood similarity, etc). This side knowledge is also shown to constrain the hypothesis space in various ways. In this section, we will attempt to answer the following statistical question: what effect do these constraints have on the generalization ability of the learned model? We will compute bounds on the complexity of the hypothesis space when the types of constraints seen in Sect. 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

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

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). Unless otherwise specified, the feature vectors in feature space $\mathcal{X}$ are bounded in norm by constant $X_b > 0$ and the coefficient vectors of the linear function class $\mathcal{F}$ are bounded in norm with constant $B_b > 0$.

3.2 Complexity measures within generalization bounds

Given these definitions, a generalization bound statement can be written as follows (Bartlett and Mendelson 2002): With probability at least $1 - \delta$ over the training sample S,

$$\forall f \in \mathcal{F}, \ \ \mathbb{E}_{x,y}[l(f(x), y)] \leq \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i) + 4\mathcal{L}\bar{\mathcal{R}}(\mathcal{F}_{|S}) + O\left(\sqrt{\frac{\log \frac{1}{\delta}}{2n}} \right),$$

where $\mathcal{L}$ is the Lipschitz constant of the loss function l. A relation between the empirical Rademacher complexity and covering number can be used to state the above uniform convergence statement in terms of the covering number. The relation (also known as Dudley's entropy integral) is (Talagrand 2005):

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) \leq c \int_0^\infty \sqrt{\frac{\log N(\sqrt{n}\epsilon, \mathcal{F}_{|S}, \|\cdot\|_2)}{n}} d\epsilon,$$

where $\mathcal{F}_{|S} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$ and c is a constant. Thus, we study upper bounds for covering numbers and empirical Rademacher complexities interchangeably through the rest of the paper.

3.3 Complexity results with a single linear constraint

We state two results: the first is based on volumetric arguments and bounds the covering number and the second is based on convex duality and bounds the empirical Rademacher complexity. The first is a result from Tulabandhula and Rudin (2014) while the second is new to this paper.

Volumetric upper bound on the covering number: Tulabandhula and Rudin (2014) analyzed the setting where a bounded linear function class is further constrained by a half space. The motivation there was to study a specific type of side knowledge, namely knowledge about the cost to solve a decision problem associated with the learning problem. The result there extends well beyond operational costs and is applicable to our setting where we have a ℓ_2 bounded linear function class with a single half space constraint.

Theorem 1 *(Theorem 2 of Tulabandhula and Rudin 2014) Let* $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_2 \leq X_b\}$ *with* $X_b > 0$*, and let* $\mu_{\mathcal{X}}$ *be the marginal probability measure on* $\mathcal{X}$*. Let*

$$\mathcal{F} = \left\{ f \,|\, f : \mathcal{X} \mapsto \mathcal{Y},\, f(x) = \beta^T x,\, \|\beta\|_2 \leq B_b,\ a^T \beta \leq 1 \right\},$$

with $B_b > 0$*. Let* $\mathcal{F}_{|S} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$*. Then for all* $\epsilon > 0$*, for any sample* S*,*

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

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

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

Intuition: 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.

Relation to standard results: It is known (Kolmogorov and Tikhomirov 1959) that set $\mathcal{B} = \{\beta : \|\beta\|_2 \leq B_b\}$ (with $B_b > 0$ being a fixed constant as before) has a bound on its covering number of the form $N(\epsilon, \mathcal{B}, \|\cdot\|_2) \leq \left(\frac{2B_b}{\epsilon} + 1\right)^p$. Since in Theorem 1 the same term appears, multiplied by a factor that is at most one and that can be substantially less than one, the bound in Theorem 1 can be tighter.

The above result bounds the covering number complexity for the hypothesis set. Next, we will bound the empirical Rademacher complexity for the same hypothesis set as above.

3.3.1 Convex duality based upper bound on empirical Rademacher complexity

Consider the setting in Theorem 1. Let $x_i \in \mathcal{X} = \{x : \|x\|_2 \leq X_b\}$ for $i = 1, ..., n$ as before. Our attempt to use convex duality to upper bound empirical Rademacher complexity yields the following bound.

Proposition 1 *Let* $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_2 \leq X_b\}$ *with* $X_b > 0$ *and*

$$\mathcal{F} = \left\{ f | f : \mathcal{X} \mapsto \mathcal{Y}, f(x) = \beta^T x, \|\beta\|_2 \leq B_b, a^T \beta \leq 1 \right\},$$

with $B_b > 0$. *Then,*

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) \leq \max \left(\mathbb{E}_\sigma \left[\min_{\eta \geq 0} (B_b \|X_L \sigma - \eta a\|_2 + \eta) \right], \mathbb{E}_\sigma \left[\min_{\eta \geq 0} (B_b \|X_L \sigma + \eta a\|_2 + \eta) \right] \right),$$

where $X_L = [x_1 \ \ldots \ x_n]$ *is a* $p \times n$ *dimensional feature matrix and* σ *is a* $n \times 1$ *dimensional vector of Bernoulli random variables taking values in* $\{-1, 1\}$.

Intuition: We can understand the effect of the linear constraint on the upper bound through the magnitude of vector a. Without loss of generality, let the expectation of the optimal value of the first minimization problem be higher (both minimization problems are structurally similar to each other except for a sign change within the norm term). For a fixed value of σ, this minimization problem involves the distance of vector $X_L\sigma$ to the scaled vector a in the first term and the scaling factor η itself as the second term.

Thus, generally, if $\|a\|_2$ is large, the scaling factor η can be small, resulting in a lower optimal value. We also know that larger $\|a\|_2$ corresponds to a tighter half space constraint. Thus, as the linear constraint on the hypothesis space becomes tighter, it makes the optimal solution η and the optimal value smaller for each σ vector. As a result, it tightens the upper bound on the empirical Rademacher complexity.

Relation to standard results: An upper bound on each term of the max operation above can be found by setting $\eta = 0$ that recovers the standard upper bound of $\frac{B_b \sqrt{\text{trace}(X_L^T X_L)}}{\sqrt{n}}$ or $\frac{B_b X_b}{\sqrt{n}}$ without capturing the effect of the linear constraint $a^T \beta \leq 1$.

3.4 Complexity results with polygonal/multiple linear constraints and general norm constraints

The following result is from Tulabandhula and Rudin (2013), where the authors analyze the effect of decision making bias on generalization of learning. Again, as in the single linear constraint case, the result extends beyond the setting considered in that paper. In particular, it covers all the motivating scenarios described in Sect. 2.1.

Let us define the matrix $[x_1 \ \ldots \ x_n]$ as matrix X_L where $x_i \in \mathcal{X} = \{x : \|x\|_r \leq X_b\}$ and $X_b > 0$. 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 $B_b > 0$ are known constants. In other words, we have V linear constraints in addition to a ℓ_q norm constraint. As before, let $\mathcal{F}_{|S}$ be the restriction of $\mathcal{F}$ with respect to S.

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

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

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: $P^K := \left\{(k_1, ..., k_p) \in \mathbb{Z}^p : \sum_{j=1}^p |k_j| \leq K\right\}$, and $P_c^K := \left\{(k_1, ..., k_p) \in P^K : \sum_{j=1}^p \tilde{c}_{j\nu} k_j \leq K \; \forall \nu = 1, ..., V\right\}$. 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. Define X_{sL} to be equal to the product of a diagonal matrix (whose j^{th} diagonal element is $\frac{n^{1/r} X_b B_b}{\|h_j\|_r}$) and X_L. Define $\lambda_{\min}(X_{sL} X_{sL}{}^T)$ to be the smallest eigenvalue of the matrix $X_{sL} X_{sL}{}^T$.

Theorem 2 *(Theorem 6 of Tulabandhula and Rudin 2013)*

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

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*

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

Intuition: The linear assumptions on the labels of the unlabeled examples $\{\tilde{x}_i\}_{i=1}^m$ determine the parameters $\{\tilde{c}_{j\nu}\}_{j,\nu}$ that in turn influence the complexity measure bound. In particular, 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. This leads to a tighter upper bound on the covering number.

Relation to standard results: We recover the covering number bound for linear function classes given in (Zhang 2002) when there are no linear constraints. In this case, the polytope P^K is well structured and the number of integer points in it can be upper bounded in an explicit way combinatorially.

It is possible to convex duality to upper bound the empirical Rademacher complexity as we did in Proposition 1. However, the intuition is less clear, and thus, we omit the bound here.

3.5 Complexity results with quadratic constraints

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

3.5.1 Upper bound on empirical Rademacher complexity

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

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

Theorem 4 *(Rademacher complexity of linear function class with two quadratic constraints) Let*

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

with A_2 symmetric positive-semidefinite and $B_b > 0$. Then,

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

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

Intuition: If the quadratic constraints are such they correspond to small ellipsoids, then the circumscribing ellipsoid will also be small. Correspondingly, the eigenvalues of $A_{\text{int}\gamma}$ will be large. Since, the upper bound depends inversely on the magnitude of these eigenvalues (since it depends on $A_{\text{int}\gamma}^{-1}$), it becomes tighter. Also, in the setting where the original ellipsoids are large and elongated but their intersection region is small and can be bounded by a small circumscribing ellipsoid, the upper bound is again tighter.

Relation to standard results: 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 $\bar{\mathcal{R}}(\mathcal{F}_{|S}) \leq \frac{1}{n}\sqrt{\sum_{j=1}^n\sum_{i=1}^p \frac{x_{ji}^2}{\lambda_i}}$ and this can be bounded by $\frac{X_b B_b}{\sqrt{n}}$ (Kakade et al. 2008) when $A_2 = \mathbf{0}$. In that case, all of the λ_i are $\frac{1}{B_b^2}$.

Remark 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:

$$\min_{\gamma\in[0,1]} \text{trace}(X_L^T(\gamma A_1 + (1-\gamma)A_2)^{-1}X_L). \tag{2}$$

This optimization problem is a univariate non-linear program.

3.5.2 Lower bound on empirical Rademacher complexity

We will now show that the dependence of the complexity on $A_{\mathrm{int}\gamma}^{-1}$ is near optimal.

Since $A_{\mathrm{int}\gamma}$ is a real symmetric matrix, let us decompose $A_{\mathrm{int}\gamma}$ into a product $P^T DP$ where D is a diagonal matrix with the eigenvalues of $A_{\mathrm{int}\gamma}$ as its entries and P is an orthogonal matrix (i.e., $P^T P = I$). Our result, which is similar in form to the upper bound of Theorem 4, is as follows.

Theorem 5

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

where

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

C is the constant in Lemma 5, P is the orthogonal matrix from the decomposition of matrix $A_{\mathrm{int}\gamma}$ defined in Theorem 4, p and $X_b > 0$ are problem constants, X_L is the matrix $[x_1 \ \ldots \ x_n]$ with examples x_i's as its columns, and n is the number of training examples.

Intuition: The lower bound is showing that the dependence on $\sqrt{\mathrm{trace}(X_L^T A_{\mathrm{int}\gamma}^{-1} X_L)}$ is tight modulo a $\log n$ factor and a factor (κ). The $\log n$ factor is essentially due to the use of the relation between Gaussian and Rademacher complexities in our proof technique. On the other hand, κ depends on the interaction between the side knowledge about the unlabeled examples (captured through matrix P) and the feature matrix X_L. If there is no interaction, that is, PX_L has zero valued rows for all $j = 1, \ldots, p$, then the lower bound on empirical Rademacher complexity becomes equal to 0. On the other hand, when there is higher interaction between $A_{\mathrm{int}\gamma}$ (or equivalently, P) and X_L, then the factor κ grows larger, tightening the lower bound on the empirical Rademacher complexity.

The dependence of the lower bound on the strength of the additional convex quadratic constraint is captured via $A_{\mathrm{int}\gamma}$ and behaves in a similar way to the upper bound. That is, when the constraint leads to a small circumscribing ellipsoid, the eigenvalues of $A_{\mathrm{int}\gamma}^{-1}$ are small and the lower bound is small (just like the upper bound). On the other hand, if the constraint leads to a larger circumscribing ellipsoid, the eigenvalues of $A_{\mathrm{int}\gamma}^{-1}$ are large, leading to a higher values of the lower bound (the upper bound also increases similarly).

Relation to standard results: As with the upper bound, when there is no second quadratic constraint, $A_{\mathrm{int}\gamma} = \frac{1}{B_b^2}\mathbb{I}$. The lower bound depends on the training data through the term $\sqrt{\mathrm{trace}(X_L^T X_L)}$ in this case.

Comparison to the upper bound: For comparison, we see that the upper bound in Theorem 4 is of the form $\frac{1}{n}\sqrt{\mathrm{trace}(X_L^T A_{\mathrm{int}\gamma}^{-1} X_L)}$ while the lower bound of Theorem 5 is of the form

$$\frac{\kappa}{n \log n}\sqrt{\mathrm{trace}(X_L^T A_{\mathrm{int}\gamma}^{-1} X_L)},$$

where κ depends on $A_{\mathrm{int}\gamma}$ and X_L.

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 the work of Wainwright (2011) for handling single ellipsoids.

3.5.3 Comparison of empirical Rademacher complexity upper bound with a covering number based bound

When matrix $A_{\text{int}\gamma}$ describing a circumscribing ellipsoid has eigenvalues $\{\lambda_i\}_{i=1}^p$, then the covering number can be bounded as:

$$N(\sqrt{n}\epsilon, \mathcal{F}_{|S}, \|\cdot\|_2) \leq \Pi_{i=1}^p \left(\frac{2X_b}{\epsilon\sqrt{\lambda_i}} + 1\right).$$

To get a tight bound, among all circumscribing ellipsoids, we should pick one that minimizes the right hand side of the bound. To do this, we solve an optimization problem involving volume minimization that is different than in (2). For instance, this volume minimization can be done using the following steps if at least one of the matrices among A_1 and A_2 is positive-definite:

- First, A_1 and A_2 are simultaneously diagonalized by congruence (say with a non-singular matrix called C) to obtain diagonal matrices $\text{Diag}(a_{1i})$ and $\text{Diag}(a_{2i})$. We can guarantee that the set of ratios $\{\frac{a_{1i}}{a_{2i}}\}$ obtained will be unique.
- The desired ellipsoid $A_{\text{int}\gamma^*}$ can then be obtained by computing

  $$\gamma^* \in \arg\max_{\gamma\in[0,1]} \Pi_{i=1}^p(\gamma a_{1i} + (1-\gamma)a_{2i})$$

  and then multiplying the optimal diagonal matrix $\text{Diag}(\gamma^* a_{1i} + (1-\gamma^*)a_{2i})$ with the congruence matrix C appropriately. Optimal γ^* can be found in polynomial time (for example, using Newton-Raphson).

Comparison with the duality approach to upper bounding empirical Rademacher complexity: A convex duality based upper bound can be derived as shown below.

Theorem 6 *Consider the setting of Theorem 4. Then,*

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) \leq \inf_{\eta\in[0,1]} \left\{\frac{1}{4n} trace(X_L^T A_{int\eta}^{-1} X_L) + \frac{1}{n}(B_b^2 + \eta(1 - B_b^2))\right\}, \tag{3}$$

where $A_{int\eta} = \mathbb{I} + \eta(A_2 - \mathbb{I})$.

This upper bound looks similar to the result in Eq. (1). Note that $A_{\text{int}\eta}$ is different from $A_{\text{int}\gamma}$ in Theorem 4. $A_{\text{int}\gamma}$ comes from a circumscribing ellipsoid, whereas $A_{\text{int}\eta}$ does not.

Instead, the matrix $A_{\text{int}\eta}$ is picked such that η minimizes the right hand side of the bound in Eq. 3. Qualitatively, we can see that if the matrix A_2 corresponding to the second ellipsoid constraint has large eigenvalues (for instance, when the second ellipsoid is a smaller sphere, or is an elongated thin ellipsoid), then $A_{\text{int}\eta}^{-1}$ is ‘small’ (the eigenvalues are small) leading to a tighter upper bound on the empirical Rademacher complexity.

3.5.4 Extension to multiple convex quadratic constraints

Although Sect. 3.5 deals with only two convex quadratic constraints, the same strategy can be used to upper bound the complexity of hypothesis class constrained by multiple convex quadratic constraints. In particular, let $\mathcal{F} = \{f : f = \beta^T x, \beta^T A_k \beta \leq 1 \ \ \forall k = 1, ..., K\}$. Again, assume one of the matrices A_k is positive definite. We can approach this problem in two stages. In the first step, we find an ellipsoid $\Xi_{\text{int}\gamma}$ (with matrix $A_{\text{int}\gamma}$) circumscribing the intersections of the K original ellipsoids and in the second step, we reuse Theorem 4 to obtain an upper bound in $\bar{\mathcal{R}}(\mathcal{F}_{|S})$.

We will generalize Eq. (2) to look for a circumscribing ellipsoid from the family of ellipsoids parameterized by a K dimensional vector γ constrained to the $K-1$ simplex. In other words, the family of circumscribing ellipsoids is given by $\{\beta^T A_{\text{int}\gamma}\beta \leq 1 : A_{\text{int}\gamma} = \sum_{k=1}^{K} \gamma_k A_k, \sum_{k=1}^{K} \gamma_k = 1, \gamma_k \geq 0 \quad \forall k = 1, ..., K\}$. We can pick one circumscribing ellipsoid from this family by minimizing the right hand side of Eq. 1 over the $K-1$ simplex similar to Eq. (2):

$$\min_{\gamma \in \left\{\gamma : \sum_{k=1}^{K} \gamma_k = 1, \gamma_k \geq 0 \;\; \forall k = 1,...,K\right\}} \text{trace}\left(X_L^T \left(\sum_{k=1}^{K} \gamma_k A_k\right)^{-1} X_L\right).$$

The above optimization problem is a $K-1$ dimensional polynomial optimization problem.

3.6 Complexity results with linear and quadratic constraints

Consider now the setting where we have both linear and quadratic constraints. In particular, we can have the assumptions leading to linear constraints and those leading to quadratic constraints hold simultaneously. In such a setting, based on Theorems 2 and 3, we can get a potentially tighter covering number result as follows. Let $x_i \in \mathcal{X} = \{x : \|x\|_2 \leq X_b\}$. Let the function class $\mathcal{F}$ be

$$\mathcal{F} = \Big\{ f | f(x) = \beta^T x, \beta \in \mathbb{R}^p, \beta^T A_1 \beta \leq 1, \beta^T A_2 \beta \leq 1, \\ \sum_{j=1}^{p} c_{j\nu}\beta_j + \delta_\nu \leq 1, \delta_\nu > 0, \nu = 1, ..., V \Big\},$$

where $\{c_{j\nu}\}_{j,\nu}$, $\{\delta_\nu\}_\nu$, A_1 and A_2 are known beforehand.

Let matrix $A_{\text{int}\gamma}$ be such that $\{\beta : \beta^T A_1 \beta \leq 1, \beta^T A_2 \beta \leq 1\}$ is circumscribed by $\{\beta : \beta^T A_{\text{int}\gamma}\beta \leq 1\}$. Defining $\{\tilde{c}_{j\nu}\}$ and X_{sL} in the same way as in Sect. 3.3, we get the following corollary.

Corollary 1 *(of Theorem 2)*

$$N(\sqrt{n}\epsilon, \mathcal{F}_{|S}, \|\cdot\|_2) \leq \begin{cases} \min\{|P^{K_0}|, |P_c^K|\} & \text{if } \epsilon < X_b\sqrt{\lambda_{\max}(A_{int\gamma}^{-1})} \\ 1 & \text{otherwise} \end{cases}.$$

Here, $K_0 = \left\lceil \frac{X_b^2 \lambda_{\max}(A_{int\gamma}^{-1})}{\epsilon^2} \right\rceil$ *and* K *is the maximum of* K_0 *and*

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

The corollary holds for any $A_{\text{int}\gamma}$ that satisfies the circumscribing requirement. In particular, we can construct the ellipsoid $\{\beta : \beta^T A_{\text{int}\gamma}\beta \leq 1\}$ such that it 'tightly' circumscribes the set $\{\beta : \beta^T A_1 \beta \leq 1, \beta^T A_2 \beta \leq 1\}$ using Theorem 3 in the same way as we did in Sect. 3.5. The intuition for how the parameters of our side knowledge, namely, the linear inequality coefficients and the matrices corresponding to the ellipsoids, is the same as in Sects. 3.4 and 3.5. Relation to standard results have also been discussed in these sections.

3.6.1 *Extension to arbitrary convex constraints*

There are at least three ways to reuse the results we have with linear, polygonal, quadratic and conic constraints to give upper bounds on covering number or empirical Rademacher complexity of function classes with arbitrary convex constraints. Such arbitrary convex constraints can arise in many settings. For instance, when the convex quadratic constraints in Sect. 2.2 are not symmetric around the origin, we cannot use the results of Sect. 3.5 directly, but the following techniques apply. Other typical convex constraints include those arising from likelihood models, entropy biases and so on.

The first approach involves constructing an outer polyhedral approximation of the convex constraint set. For instance, if we are given a separation oracle for the convex constraint, constructing an outer polyhedral approximation is relatively straightforward. We can also optimize for properties like the number of facets or vertices of the polyhedron during such a construction. Given such an outer approximation, we can apply Theorem 2 to get an upper bound on the covering number of the hypothesis space with the given convex constraint.

The second approach involves constructing a circumscribing ellipsoid for the constraint set. This is possible for any convex set in general (John 1948). In addition if the convex set is symmetric around the origin, the ‘tightness’ of the circumscribing ellipsoid improves by a factor $\sqrt{p}$, where p is the dimension of the linear coefficient vector β. Given such a circumscribing ellipsoid, we can apply Theorem 4 to get an upper bound on the empirical Rademacher complexity of the original function class with the convex constraint. The quality of both of these outer relaxation approaches depends on the structure and form of the convex constraint we are given.

The third approach is to analyze the empirical Rademacher complexity directly using convex duality as we have done for the linear and quadratic cases, and as we will do for the conic case next.

3.7 Complexity results with multiple conic constraints

Consider the function class

$$\mathcal{F} = \{f : f = \beta^T x, \beta^T\beta \le B_b^2, \|A_k\beta\|_2 \le a_k^T\beta + d_k \quad \forall k = 1, \ldots, K\},$$

where we have one convex quadratic constraint and K conic constraints. We can find an upper bound on the Rademacher complexity as shown below.

Theorem 7 *(Rademacher complexity of bounded linear function class with conic constraints) Let $\mathcal{X} = \{x : \|x\|_2 \le X_b\}$ with $X_b > 0$ and let*

$$\mathcal{F} = \{f : f = \beta^T x, \beta^T\beta \le B_b^2, \|A_k\beta\|_2 \le a_k^T\beta + d_k \quad \forall k = 1, \ldots, K\},$$

where $B_b > 0$,$\{A_k, a_k, d_k\}_{k=1}^K$ are the parameters. Assume $A_k \succ 0$ and let $\lambda_{\min}(A_k)$ denote its minimum eigenvalue for $k = 1, \ldots, K$. Also let $\sup_{x\in\mathcal{X}} \|x\|_2 \le X_b$. Then,

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) \le \frac{X_b}{\sqrt{n}} \cdot \min\left\{B_b, \sum_{k=1}^{K} \frac{B_b\|a_k\|_2 + d_k}{K \cdot \lambda_{\min}(A_k)}\right\}.$$

Intuition: When $\|a_k\|_2$ and d_k are $o(\lambda_{\min}(A_k))$, the effect of conic constraints can influence the upper bound on the empirical Rademacher complexity and make the corresponding generalization bounds tighter. From a geometric point of view, we can infer the following:

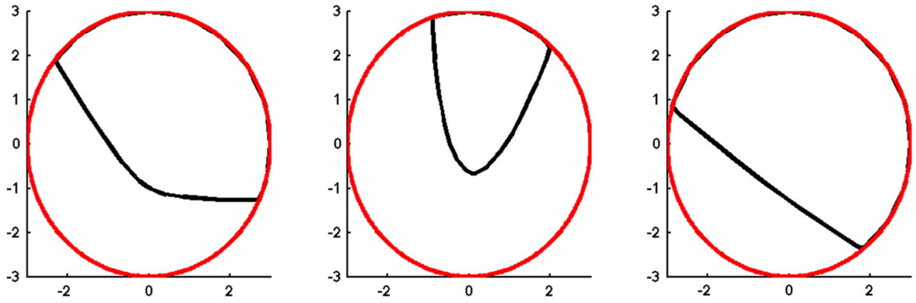

Fig. 2 Here we illustrate the effect of a single conic constraint $\{\beta : \sqrt{4\mu\beta_1^2 + \mu\beta_2^2} \leq \delta(2\beta_1 + 3\beta_2 + 4)\}$ on our hypothesis space $\{\beta \in \mathbb{R}^2 : \beta^T\beta \leq 9\}$ for different scaling values of parameters μ and δ. In our notation, matrix $A = [2\sqrt{\mu} \;\; 0; 0 \;\; \sqrt{\mu}]$, vector $a = \delta[2 \;\; 3]^T$ and scalar $d = 4\delta$. *Left:* Parameter set (μ, δ) is equal to $(1, 1)$. The region covered by the conic constraint is the convex set in the upper part of the circle. *Center:* Changing the parameters (μ, δ) to $(10, 1)$ makes the eigenvalue $\lambda_{\min}(A)$ larger thus reducing the intersection region further. *Right:* Changing the parameters (μ, δ) to $(1, 10)$ increases the magnitude of $\|a\|_2$ and d relative to the value of $\lambda_{\min}(A)$ increases the intersection region between the conic constraint and the ball. This leads to a larger empirical Rademacher complexity bound value

if the cones are sharp, then $\lambda_{\min}(A_k)$ are high, implying a smaller empirical Rademacher complexity. Figure 2 illustrates this in two dimensions.

Relation to standard results: The looser unconstrained version of the upper bound $\frac{X_b B_b}{\sqrt{n}}$ is recovered when there are no conic constraints or when the conic constraints are ineffective (for instance, when $\|a_k\|_2$ is high, d_k is a large offset or $\lambda_{\min}(A_k)$ is small).

Remark 3 There have been some recent attempts to obtain bounds on a related measure, similar to the empirical Gaussian complexity defined here, in the compressed sensing literature that also involves conic constraints (Stojnic 2009). Their objective (minimum number of measurements for signal recovery assuming sparsity) is very different from our objective (function class complexity and generalization). In the former context, there are a few results (Chandrasekaran et al. 2012) dealing with the intersection of a single generic cone with a sphere ($\mathbb{S}^{p-1}$) whereas in this context, we look at the intersection of multiple second order cones (explicitly parameterized by $\{A_k, a_k, d_k\}_{k=1}^K$) with balls ($\{\beta^T\beta \leq B_b^2\}$).

4 Related work

It is well-known that having additional unlabeled examples can aid in learning (Shental et al. 2004; Nguyen and Caruana 2008b; Gómez-Chova et al. 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 et al. 2008; Rigollet 2007) and manifold assumptions (Belkin and Niyogi 2004; Belkin et al. 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.

Researchers have also attempted to incorporate domain knowledge directly into learning algorithms, where this domain knowledge does not necessarily arise from unlabeled examples. For instance, the framework of knowledge based SVMs (Fung et al. 2002; Le et al. 2006) motivates the use of various constraints or modifications in the learning procedure to incorporate specific kinds of knowledge (without using unlabeled examples). The focus of Fung et al. (2002) is algorithmic and they consider linear constraints. Le et al. (2006) incorporate knowledge by modifying the function class itself, for instance, from linear function to non-linear functions.

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. Specifically, in the work of Balcan and Blum (2005), the authors introduce the notion of incompatibility $\mathbb{E}_{x\sim D}[1-\chi(h,x)]$ between a function h and the input distribution D. The unlabeled examples are used to estimate the distribution dependent quantity $\mathbb{E}_{x\sim D}[1-\chi(h,x)]$. By imposing the constraint that models have their incompatibility with the distribution of the data source D below a desired level, we restrict the hypothesis space. Their result for a finite hypothesis space is as follows:

Theorem 8 *(Theorem 1 of Balcan and Blum 2005) If we see m unlabeled examples and n labeled examples, where*

$$m \geq \frac{1}{\epsilon}\left[\ln|C| + \ln\frac{2}{\delta}\right] \quad \textit{and} \quad n \geq \frac{1}{\epsilon}\left[\ln|C_{D,\chi}(\epsilon)| + \ln\frac{2}{\delta}\right],$$

then with probability $1-\delta$, *all* $h \in C$ *with zero training error and zero incompatibility* $\frac{1}{m}\sum_{i=1}^{m}(1-\chi(h,\tilde{x}_i)) = 0$, *we have* $\mathbb{E}[l(h(x),y)] \leq \epsilon$.

Here C is the finite hypothesis space of which h is an element and $C_{D,\chi}(\epsilon) = \{h \in C : \mathbb{E}_{x\sim D}[1-\chi(h,x)] \leq \epsilon\}$. In the work of Kääriäinen (2005), the author obtains a generalization bound by approximating the disagreement probability of pairs of classifiers using unlabeled data. Again, here the unlabeled data is used to estimate a distribution dependent quantity, namely, the true disagreement probability between consistent models. In particular, the disagreement between two models h and g is defined to be $d(h,g) = \frac{1}{m}\sum_{i=1}^{m} 1_{[h(\tilde{x}_i)\neq g(\tilde{x}_i)]}$. The following theorem about generalization is proposed.

Theorem 9 *Let* $\mathcal{F}$ *be the class of consistent models, that is, the set of models with zero training error. Assume the true model belongs to this class. Let* $\hat{f} \in \mathcal{F}$ *be the function whose distance to the farthest function in* $\mathcal{F}$ *is minimal (via metric d). Then, for all S, with probability* $1-\delta$ *over the choice of unlabeled sample* $S^{unlabeled}$,

$$\mathbb{E}_{S^{unlabeled}}[l(\hat{f}(x),y)] \leq \inf_{f\in\mathcal{F}} \sup_{g\in\mathcal{F}} d(f,g)$$
$$+\bar{\mathcal{R}}(\{1_{[g\neq g']}|g,g' \in F\}_{|S^{unlabeled}}) + O\left(\sqrt{\frac{\ln(2/\delta)}{m}}\right).$$

Note that the randomization in both Theorems 8 and 9 is also over unlabeled data. In our theorems, we do not randomize with respect to the unlabeled data. For us, they serve a different purpose and do not need to be chosen randomly. 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 Proofs

5.1 Proof of Proposition 1

Proof Instead of working with the maximization problem in the definition of empirical Rademacher complexity, we will work with a couple of related maximization problems, due to the following lemma.

Lemma 1

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) \leq \mathbb{E}\left[\max\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i),\ \sup_{f\in\mathcal{F}} -\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right)\right]. \tag{4}$$

Proof Since the empirical Rademacher complexity is defined as $\mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} \frac{1}{n}|\sum_{i=1}^{n}\sigma_i f(x_i)|]$, we will show that for any fixed σ vector,

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i f(x_i)\right| \leq \max\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i),\ \sup_{f\in\mathcal{F}} -\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right). \tag{5}$$

The inequality above is straightforward to prove. Let f^* be the optimal solution to the maximization problem on the left. Then, f^* is a feasible point for each of the maximization problems on the right. We will look at two cases: In the first case, let $\frac{1}{n}\sum_{i=1}^{n}\sigma_i f^*(x_i) \geq 0$. Then, clearly the first maximization problem on the right, namely, $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)$ will have an optimal value greater than or equal to the left side of Eq. (5). In the second case, let $\frac{1}{n}\sum_{i=1}^{n}\sigma_i f^*(x_i) < 0$. Then, the second maximization problem on the right, namely, $\sup_{f\in\mathcal{F}} -\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)$ will have an optimal value greater than or equal to the left side of Eq. (5). That is, in this case:

$$0 \leq \left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f^*(x_i)\right| = -\frac{1}{n}\sum_{i=1}^{n}\sigma_i f^*(x_i) \leq \sup_{f\in\mathcal{F}} -\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i).$$

Combining the two cases, we get the Eq. (5). Taking expectations over σ gives us the desired inequality.

Continuing with the proof of Proposition 1: Let $g = \sum_{i=1}^{n}\sigma_i x_i = X_L\sigma$ so that $\bar{\mathcal{R}}(\mathcal{F}_{|S}) = \frac{1}{n}\mathbb{E}[\sup_{\beta\in\mathcal{F}} |g^T\beta|]$. We will attempt to dualize the two maximization problems in the upper bound provided by Lemma 1 to get a bound on the empirical Rademacher complexity. Both maximization problems are very similar except for the objective. Let $\omega(g, \mathcal{F})$ be the optimal value of the following optimization problem:

$$\begin{aligned} \max_{\beta}\ g^T\beta \quad \text{s.t.} \\ \beta^T\beta \leq B_b^2 \\ a^T\beta \leq 1. \end{aligned}$$

Thus $\omega(g, \mathcal{F})$ represents the optimal value of the maximization problem inside the expectation operation in the first term of Eq. (4). We will now write a dual program to the above and use weak duality to upper bound $\omega(g, \mathcal{F})$. The Lagrangian is:

$$\mathcal{L}(\beta, \gamma, \eta) = g^T \beta + \gamma(B_b^2 - \beta^T \beta) + \eta(1 - a^T \beta),$$

where $\beta \in \mathbb{R}^p$, $\gamma \in \mathbb{R}_+$, $\eta \in \mathbb{R}_+$. Maximizing the Lagrangian with respect to β gives us:

$$\begin{aligned}
&\max_\beta \ \mathcal{L}(\beta, \gamma, \eta) \\
&\quad = \max_\beta \left[(g - \eta a)^T \beta - \gamma \beta^T \beta + \gamma B_b^2 + \eta\right] \\
&\quad = \max_\beta \left[-\gamma \left[\beta^T \beta - \frac{2(g - \eta a)^T \beta}{2\gamma} + \frac{\|g - \eta a\|_2^2}{4\gamma^2}\right] + \frac{\|g - \eta a\|_2^2}{4\gamma} + \gamma B_b^2 + \eta\right] \\
&\quad = \max_\beta \left[-\gamma \left\|\beta - \frac{g - \eta a}{2\gamma}\right\|_2^2 + \frac{\|g - \eta a\|_2^2}{4\gamma} + \gamma B_b^2 + \eta\right] \\
&\quad = \frac{\|g - \eta a\|_2^2}{4\gamma} + \gamma B_b^2 + \eta.
\end{aligned}$$

The dual problem is thus

$$\min_{\gamma \geq 0, \eta \geq 0} \frac{\|g - \eta a\|_2^2}{4\gamma} + \gamma B_b^2 + \eta.$$

Minimizing with respect to one of the decision variables, γ, gives the following dual problem

$$\min_{\eta \geq 0} B_b \|g - \eta a\|_2 + \eta.$$

Thus, $\omega(g, \mathcal{F}) \leq \min_{\eta \geq 0}(B_b \|g - \eta a\|_2 + \eta)$. Similarly we can prove an upper bound on the maximization problem appearing in the second term in the max operation in Eq. (4), which will be $\min_{\eta \geq 0}(B_b \|g + \eta a\|_2 + \eta)$. Thus, the empirical Rademacher complexity is upper bounded as:

$$\begin{aligned}
\bar{\mathcal{R}}(\mathcal{F}_{|S}) &\leq \frac{1}{n} \max\left(\mathbb{E}\left[\min_{\eta \geq 0}(B_b \|g - \eta a\|_2 + \eta)\right], \mathbb{E}\left[\min_{\eta \geq 0}(B_b \|g + \eta a\|_2 + \eta)\right]\right) \\
&= \frac{1}{n} \max\left(\mathbb{E}_\sigma\left[\min_{\eta \geq 0}(B_b \|X_L \sigma - \eta a\|_2 + \eta)\right], \mathbb{E}_\sigma\left[\min_{\eta \geq 0}(B_b \|X_L \sigma + \eta a\|_2 + \eta)\right]\right).
\end{aligned}$$

□

5.2 Proof of Theorem 4

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

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

$$
\begin{aligned}
&= \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{\{\alpha:\|\alpha\|_2\le 1\}}\left|\alpha^T(A_{\text{int}\gamma}^{-1/2})^T X_L\sigma\right|\right]\\
&\overset{(c)}{=} \frac{1}{n}\mathbb{E}_\sigma\left[\|(A_{\text{int}\gamma}^{-1/2})^T X_L\sigma\|_2\right]\\
&\overset{(d)}{\le} \frac{1}{n}\sqrt{\mathbb{E}_\sigma\left[\|(A_{\text{int}\gamma}^{-1/2})^T X_L\sigma\|_2^2\right]}\\
&= \frac{1}{n}\sqrt{\mathbb{E}_\sigma\left[\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L\sigma\sigma^T)\right]}\\
&\overset{(e)}{=} \frac{1}{n}\sqrt{\text{trace}(X_L^T A_{\text{int}\gamma}^{-1} X_L)}
\end{aligned}
$$

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 $\mathbb{E}[\sigma\sigma^T] = I$. □

5.3 Proof of Theorem 5

Proof Recall that we can decompose $A_{\text{int}\gamma}$ into a product $P^T DP$ 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., $P^T P = I$). Let us define a new variable: $\alpha := P\beta$, which is a linear transformation of linear model parameter β. Then, the empirical Gaussian complexity of our function class can be written as:

$$\bar{\mathcal{G}}(\mathcal{F}_{|S}) = \mathbb{E}_\sigma\left[\sup_{\alpha^T D\alpha\le 1}\frac{1}{n}\sum_{i=1}^{n}\left|\sigma_i\alpha^T Px_i\right|\right],$$

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 $\sum_{i=1}^n \sigma_i x_i$. That is, let $\omega(\sigma) := P\sum_{i=1}^n \sigma_i x_i$. We will drop the dependence of ω on σ from the notation when it is clear from the context. The expression now becomes

$$n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S}) \ge \mathbb{E}_\sigma\left[\sup_{\alpha^T D\alpha\le 1}\alpha^T\omega\right], \tag{6}$$

where the inequality is because we removed the absolute sign in the right hand side expression before substituting for ω.

The following are the major steps in our proof:

- We will analyze the Gaussian function $F(\omega(\sigma)) := \sup_{\alpha^T D\alpha\le 1}\alpha^T\omega(\sigma)$ and show it is Lipschitz in σ. This is proved in Lemma 2.
- Then we apply Lemma 3, which is about Gaussian function concentration, to the above function. In particular, we will upper bound the variance of the Gaussian function $F(\omega(\sigma))$ in terms of its parameters (Lipschitz constant, matrix D, etc).
- We then generate a candidate lower bound for the empirical Gaussian complexity.
- The upper bound on the variance of $F(\omega(\sigma))$ we found earlier is used to make this bound proportional to $\sqrt{\text{trace}(X_L A_{\text{int}\gamma}^{-1} X_L)}$.
- Finally, we use a relation between empirical Rademacher complexity and empirical Gaussian complexity to obtain the desired result.

5.3.1 Computing a Lipschitz constant for $F(\omega(\sigma))$

The following lemma gives an upper bound on the Lipschitz constant of $F(\omega(\sigma))$.

Lemma 2 *The function $F(\omega(\sigma)) := \sup_{\alpha^T D\alpha \le 1} \alpha^T \omega(\sigma)$ is Lipschitz in σ with a Lipschitz constant $\mathcal{L}$ bounded by $X_b\sqrt{\frac{p\cdot n}{\lambda_{min}(D)}}$.*

Proof We have

$$F(\omega) = \sup_{\alpha^T D\alpha \le 1} \alpha^T \omega = \sup_{(D^{1/2}\alpha)^T(D^{1/2}\alpha)\le 1} \alpha^T \omega.$$

Using a new dummy variable $\rho = D^{1/2}\alpha$ we have:

$$F(\omega) = \sup_{\rho^T\rho\le 1} (D^{-1/2}\rho)^T \omega = \sup_{\rho^T\rho\le 1} \rho^T (D^{-1/2})^T \omega = \|D^{-1/2}\omega\|_2.$$

Thus,

$$\begin{aligned}|F(\omega_1) - F(\omega_2)| &= \left|\|D^{-1/2}\omega_1\|_2 - \|D^{-1/2}\omega_2\|_2\right| \le \|D^{-1/2}(\omega_1 - \omega_2)\|_2 \\ &\overset{(a)}{\le} \left\|\frac{1}{\sqrt{\lambda_{min}(D)}} I(\omega_1 - \omega_2)\right\|_2 = \frac{1}{\sqrt{\lambda_{min}(D)}}\|\omega_1 - \omega_2\|_2.\end{aligned}$$

At (a), we used the fact that $D^{-1} \preceq \frac{1}{\lambda_{min}(D)} I$.

Now, we will upper bound $\|\omega_1 - \omega_2\|_2$ using σ_1 and σ_2 as follows. Using the definition of $\omega = PX_L\sigma$ we get,

$$\begin{aligned}\|\omega_1 - \omega_2\|_2 &= \|PX_L\sigma_1 - PX_L\sigma_2\|_2 = \|PX_L(\sigma_1 - \sigma_2)\|_2 \\ &\overset{(b)}{\le} \|X_L(\sigma_1 - \sigma_2)\|_2 \\ &= \sqrt{(\sigma_1 - \sigma_2)^T X_L^T X_L(\sigma_1 - \sigma_2)} \\ &\overset{(c)}{\le} \sqrt{(\sigma_1 - \sigma_2)^T \lambda_{max}(X_L^T X_L) I(\sigma_1 - \sigma_2)} \\ &= \sqrt{\lambda_{max}(X_L^T X_L)}\|\sigma_1 - \sigma_2\|_2 \\ &\overset{(d)}{\le} X_b\sqrt{p\cdot n}\|(\sigma_1 - \sigma_2)\|_2.\end{aligned}$$

Here, (b) follows because P is an orthonormal matrix, (c) because $X_L^T X_L \preceq \lambda_{max}(X_L^T X_L) I$ and (d) because $\lambda_{max}(X_L^T X_L) \le \text{trace}(X_L^T X_L) = \sum_{i=1}^n (X_L^T X_L)_{ii}$. Since, each diagonal element of $X_L^T X_L$ is a sum of p terms each upper bounded by X_b^2, we have $\lambda_{max}(X_L^T X_L) \le n \cdot p \cdot X_b^2$. □

Upper bounding the variance of $F(\omega(\sigma))$ using Gaussian concentration: The following lemma describes concentration for Lipschitz functions of gaussian random variables.

Lemma 3 *(Concentration, Tsirelson et al. 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*

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

The proof of Lemma 3 is omitted here. Using Lemmas 2 and 3 with $G(\sigma) = F(\omega)$, we have

$$\mathbb{P}[|(F(\omega) - \mathbb{E}_\sigma[F(\omega)]| \geq t] \leq 2e^{-\frac{t^2}{2\mathcal{L}^2}}, \tag{7}$$

where $\mathcal{L} = X_b\sqrt{\frac{p \cdot n}{\lambda_{min}(D)}}$.

Let $Y = |(F(\omega) - \mathbb{E}_\sigma[F(\omega)]|$. Then from the above tail bound, $P(Y^2 \geq s) \leq 2e^{-\frac{s}{2\mathcal{L}^2}}$ is true. Now we can bound the variance of $F(\omega)$ using the above inequality and the following lemma.

Lemma 4 *For a random variable* Y^2, $\mathbb{E}[Y^2] = \int_0^{+\infty} P(Y^2 \geq s)ds$.

Proof This is an alternate expression for the expectation of a non-negative univariate random variable in terms of its distribution function. To show this, let us assume that the density function of Y^2 is μ_{Y^2}. We then have $P(Y^2 \geq s) = 1 - P(Y^2 \leq s) = 1 - \int_0^s \mu_{Y^2}(s')ds'$ and thus: $\mu_{Y^2}(s) = -\frac{dP(Y^2 \geq s)}{ds}$. So,

$$\begin{aligned}\mathbb{E}[Y^2] &= \int_0^{+\infty} s\mu_{Y^2}(s)ds = -\int_0^{+\infty} s\frac{dP(Y^2 \geq s)}{ds}ds \\ &= -[sP(Y^2 \geq s)]_0^{+\infty} + \int_0^{+\infty} P(Y^2 \geq s)ds.\end{aligned}$$

The first term is zero and we obtain our expression. □

The variance of $F(\omega)$, which is the same as the expectation of Y^2, can thus be upper bounded as follows:

$$\begin{aligned}\mathrm{Var}(F(\omega)) &= \mathbb{E}_\sigma(Y^2) \overset{(a)}{=} \int_0^{+\infty} P(Y^2 \geq s)ds \\ &\overset{(b)}{\leq} 2\int_0^{+\infty} e^{-\frac{s}{2\mathcal{L}^2}}ds = 4X_b^2\frac{p \cdot n}{\lambda_{min}(D)},\end{aligned} \tag{8}$$

where we used Lemma 4 for step (a) and Eq. (7) for step (b) and finally substituting $X_b\sqrt{\frac{p \cdot n}{\lambda_{min}(D)}}$ for $\mathcal{L}$.

5.3.2 Lower bounding the empirical Gaussian complexity

Now we will lower bound the empirical Gaussian complexity by constructing a feasible candidate α' to substitute for the sup operation in Eq. (6). Later, we will use the variance upper bound on $F(\omega)$ we found in the earlier section to make the bound more specific.

Let $j^* \in \{1, ..., p\}$ be the index at which the diagonal element $D(j^*, j^*) = \lambda_{min}(D)$. For each realization of σ (or equivalently ω) let $\alpha' = \left[0 \ldots \frac{|\omega_{j^*}|}{\omega_{j^*}\sqrt{\lambda_{min}(D)}} \ldots 0\right]$ with the non-zero entry at coordinate j^*. Clearly α' is a feasible vector in the ellipsoidal constraint $\{\alpha : \alpha^T D\alpha \leq 1\}$ seen in the complexity expression, Eq. (6). Substituting it and using the definition of $F(\omega)$, we get a lower bound on the empirical Gaussian complexity:

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

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'.

5.3.3 Making the lower bound more specific using variance of $F(\omega(\sigma))$

Note that compared to the upper bound on the related Rademacher complexity obtained in Theorem 4, the dependence of empirical Gaussian complexity 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 Eq. (1). Rearranging the terms in the previous inequality, we get:

$$\frac{(\mathbb{E}_\sigma[F(\omega)])^2}{(\mathbb{E}_\sigma|\omega_{j^*}|)^2} \geq \frac{1}{\lambda_{min}(D)}. \tag{9}$$

By rewriting the variance in terms of the second and first moments, using expression (8) and then using (9) we get

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

Using expression (6) again, and then rearranging the terms in the previous expression, we obtain another lower bound on the scaled Gaussian complexity, which is:

$$\begin{aligned}\left(n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S})\right)^2 \geq (\mathbb{E}_\sigma[F(\omega)])^2 &\geq \frac{\mathbb{E}_\sigma[(F(\omega))^2]}{1+\frac{4pnX_b^2}{(\mathbb{E}_\sigma|\omega_{j^*}|)^2}}\\ &= \frac{\mathbb{E}_\sigma[(\sup_{\alpha^T D\alpha\leq 1}\omega^T\alpha)^2]}{1+\frac{4pnX_b^2}{(\mathbb{E}_\sigma|\omega_{j^*}|)^2}}.\end{aligned} \tag{10}$$

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 the empirical 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

$$\begin{aligned}\mathbb{E}_\sigma[|w_{j^*}|] &= \sqrt{\frac{2}{\pi}}\left(\sum_i (Px_i)^2_{j^*}\right)^{\frac{1}{2}}\\ &\geq \sqrt{\frac{2}{\pi}}\min_{j=1,\ldots,p}\|(PX_L)_j\|_2\end{aligned} \tag{11}$$

where $(PX_L)_j$ represents the jth row of the matrix PX_L. For the second moment term of (10) 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,

$$\begin{aligned}\mathbb{E}_\sigma\Big[\Big(\sup_{\alpha^T D\alpha\le 1}\omega^T\alpha\Big)^2\Big] &= \mathbb{E}_\sigma[\|D^{-1/2}PX_L\sigma\|_2^2]\\ &= \mathbb{E}_\sigma[(D^{-1/2}PX_L\sigma)^T D^{-1/2}PX_L\sigma]\\ &= \mathbb{E}_\sigma[\sigma^T X_L^T A_{\mathrm{int}\gamma}^{-1}X_L\sigma]\\ &= \mathbb{E}_\sigma[\mathrm{trace}(\sigma^T X_L^T A_{\mathrm{int}\gamma}^{-1}X_L\sigma)]\\ &= \mathbb{E}_\sigma[\mathrm{trace}(X_L^T A_{\mathrm{int}\gamma}^{-1}X_L\sigma\sigma^T)]\\ &= \mathrm{trace}(X_L^T A_{\mathrm{int}\gamma}^{-1}X_L). \qquad (12)\end{aligned}$$

Substituting the two bounds we just derived, (11) and (12), into (10) gives us a lower bound on the scaled Gaussian complexity:

$$\begin{aligned}\big(n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S})\big)^2 &\ge \frac{\mathrm{trace}(X_L^T A_{\mathrm{int}\gamma}^{-1}X_L)}{1+\frac{4pnX_b^2}{(\sqrt{\frac{2}{\pi}}\min_{j=1,\dots,p}\|(PX_L)_j\|_2)^2}}\\ n\cdot\bar{\mathcal{G}}(\mathcal{F}_{|S}) &\ge \sqrt{\frac{\mathrm{trace}(X_L^T A_{\mathrm{int}\gamma}^{-1}X_L)}{1+\frac{4pnX_b^2}{(\sqrt{\frac{2}{\pi}}\min_{j=1,\dots,p}\|(PX_L)_j\|_2)^2}}}.\end{aligned}$$

5.3.4 Using the relation between Rademacher and Gaussian complexities

The empirical Gaussian complexity is related to the empirical Rademacher complexity as follows.

Lemma 5 (Lemma 4 of Bartlett and Mendelson 2002) *There are absolute constants C and C' such that for every $\mathcal{F}_{|S}$ with $|S|=n$,*

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

Using the above result gives:

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

Thus, we get our desired result:

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

where

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

□

5.4 Proof of Corollary 1

Proof Since the ellipsoid defined using $A_{\text{int}\gamma}$ circumscribes the region of intersection of ellipsoids determined by A_1 and A_2, we have

$$\begin{aligned}\mathcal{F} = \Big\{ f \,|\, f(x) = \beta^T x, \beta \in \mathbb{R}^p, \beta^T A_1 \beta \leq 1, \beta^T A_2 \beta \leq 1,& \\ \sum_{j=1}^{p} c_{j\nu}\beta_j + \delta_\nu \leq 1, \delta_\nu > 0, \nu = 1, ..., V \Big\}& \\ \subseteq \qquad\qquad& \\ \Big\{ f \,|\, f(x) = \beta^T x, \beta \in \mathbb{R}^p, \beta^T A_{\text{int}\gamma} \beta \leq 1,& \\ \sum_{j=1}^{p} c_{j\nu}\beta_j + \delta_\nu \leq 1, \delta_\nu > 0, \nu = 1, ..., V \Big\} =: \mathcal{F}'.&\end{aligned}$$

Further, $\beta^T \lambda_{\min}(A_{\text{int}\gamma}) I \beta \leq \beta^T A_{\text{int}\gamma} \beta \leq 1$ since $\lambda_{\min}(A_{\text{int}\gamma}) I \preceq A_{\text{int}\gamma}$. That is, the set $\beta^T \lambda_{\min}(A_{\text{int}\gamma}) I \beta \leq 1$ is bigger than the ellipsoid defined using $A_{\text{int}\gamma}$. Thus,

$$\begin{aligned}\mathcal{F}' = \Big\{ f \,|\, f(x) = \beta^T x, \beta \in \mathbb{R}^p, \beta^T A_{\text{int}\gamma} \beta \leq 1,& \\ \sum_{j=1}^{p} c_{j\nu}\beta_j + \delta_\nu \leq 1, \delta_\nu > 0, \nu = 1, ..., V \Big\}& \\ \subseteq \qquad\qquad& \\ \Big\{ f \,|\, f(x) = \beta^T x, \beta \in \mathbb{R}^p, \beta^T \beta \leq \frac{1}{\lambda_{\min}(A_{\text{int}\gamma})},& \\ \sum_{j=1}^{p} c_{j\nu}\beta_j + \delta_\nu \leq 1, \delta_\nu > 0, \nu = 1, ..., V \Big\} =: \mathcal{F}''.&\end{aligned}$$

Noting that $\beta^T \beta \leq \frac{1}{\lambda_{\min}(A_{\text{int}\gamma})}$ is the same as $\|\beta\|_2 \leq \sqrt{\lambda_{\max}(A_{\text{int}\gamma}^{-1})}$, we can use Theorem 2 on $\mathcal{F}''$ with $r = 2$, $q = 2$ and $\mathcal{B}_b := \sqrt{\lambda_{\max}(A_{\text{int}\gamma}^{-1})}$ to get a bound on $N(\sqrt{n}\epsilon, \mathcal{F}''_{|S}, \|\cdot\|_2) \geq N(\sqrt{n}\epsilon, \mathcal{F}_{|S}, \|\cdot\|_2)$ giving us the stated result. □

5.5 Proof of Theorem 6

Proof Let $g = \sum_{i=1}^{n} \sigma_i x_i = X_L \sigma$ so that $\bar{\mathcal{R}}(\mathcal{F}_{|S}) = \frac{1}{n}\mathbb{E}[\sup_{\beta\in\mathcal{F}} |g^T \beta|]$. Instead of directly working with the empirical Rademacher complexity, we will dualize the two maximization problems in the upper bound given by Eq. (4) of Lemma 1. Both maximization problems are very similar except for the objective. Let $\omega(g, \mathcal{F})$ be the optimal value of the following optimization problem:

$$\begin{aligned}\max_{\beta}\, g^T \beta \quad \text{s.t.}& \\ \beta^T \beta \leq B_b^2& \\ \beta^T A_2 \beta \leq 1.&\end{aligned}$$

Thus $\omega(g, \mathcal{F})$ is proportional to the first term inside the max operation in Eq. (4), which gives an upper bound in the empirical Rademacher complexity. We will now write a dual program to the above and use weak duality to upper bound $\omega(g, \mathcal{F})$. The Lagrangian is:

$$\mathcal{L}(\beta, \gamma, \eta) = g^T\beta + \gamma(B_b^2 - \beta^T\beta) + \eta(1 - \beta^T A_2\beta),$$

where $\beta \in \mathbb{R}^p$, $\gamma \in \mathbb{R}_+$, $\eta \in \mathbb{R}_+$. Maximizing the Lagrangian with respect to β gives us:

$$\begin{aligned}
&\max_\beta \ \mathcal{L}(\beta, \gamma, \eta) = \\
&= \max_\beta \left[g^T\beta - \gamma\beta^T\beta - \eta\beta^T A_2\beta + \gamma B_b^2 + \eta \right] \\
&= \max_\beta \left[- \left(-g^T\beta + \beta^T(\gamma\mathbb{I} + \eta A_2)\beta \right) + \gamma B_b^2 + \eta \right] \\
&= \max_\beta \Big[- \Big(-g^T(\gamma\mathbb{I} + \eta A_2)^{-1/2}(\gamma\mathbb{I} + \eta A_2)^{1/2}\beta \\
&\quad +\beta^T(\gamma\mathbb{I} + \eta A_2)^{1/2}(\gamma\mathbb{I} + \eta A_2)^{1/2}\beta \Big) + \gamma B_b^2 + \eta \Big] \\
&= \max_\beta \Bigg[- \left\| (\gamma\mathbb{I} + \eta A_2)^{1/2}\beta - \frac{(\gamma\mathbb{I} + \eta A_2)^{-1/2} g}{2} \right\|_2^2 \\
&\quad + \frac{\|(\gamma\mathbb{I} + \eta A_2)^{-1/2} g\|_2^2}{4} + \gamma B_b^2 + \eta \Bigg] \\
&= \frac{\|(\gamma\mathbb{I} + \eta A_2)^{-1/2} g\|_2^2}{4} + \gamma B_b^2 + \eta,
\end{aligned}$$

where in the last step we set $\beta = \frac{(\gamma\mathbb{I}+\eta A_2)^{-1} g}{2}$. The dual problem is thus:

$$\min_{\gamma\geq 0, \eta\geq 0} \frac{\|(\gamma\mathbb{I} + \eta A_2)^{-1/2} g\|_2^2}{4} + \gamma B_b^2 + \eta \text{ , or equivalently,}$$
$$\min_{\gamma\geq 0, \eta\geq 0} \frac{1}{4} g^T(\gamma\mathbb{I} + \eta A_2)^{-1} g + \gamma B_b^2 + \eta.$$

If we let $\gamma = 1 - \eta$, we are further constraining the minimization problem, yielding another upper bound of the form:

$$\omega(g, \mathcal{F}) \leq \min_{\eta\in[0,1]} \frac{1}{4} g^T(\mathbb{I} + \eta(A_2 - \mathbb{I}))^{-1} g + B_b^2 + \eta(1 - B_b^2).$$

If we consider the second maximization problem $\sup_{\beta\in\mathcal{F}} -g^T\beta$ that appears in Eq. (4), we can similarly upper bound its optimal value with the same minimization problem as $\omega(g, \mathcal{F})$. One intuitive reason why the same minimization problem serves as an upper bound is because the hypothesis class $\mathcal{F}$ is closed under negation. Thus, we get an upper bound on the empirical Rademacher complexity as:

$$\begin{aligned}
\bar{\mathcal{R}}(\mathcal{F}_{|S}) &\leq \mathbb{E}\left[\frac{1}{n}\omega(g, \mathcal{F}) \right] \\
&\leq \mathbb{E}\left[\frac{1}{n} \min_{\eta\in[0,1]} \frac{1}{4} g^T(\mathbb{I} + \eta(A_2 - \mathbb{I}))^{-1} g + B_b^2 + \eta(1 - B_b^2) \right],
\end{aligned}$$

where recall that $g = \sum_{i=1}^{n} \sigma_i x_i$. Fix any feasible η. Let $A_{\text{int}\eta} := (\mathbb{I} + \eta(A_2 - \mathbb{I}))$ (it corresponds to an ellipsoid as well since $\eta \in [0, 1]$). Then,

$$\begin{aligned}\bar{\mathcal{R}}(\mathcal{F}_{|S}) &\le \mathbb{E}\left[\frac{1}{4n}\sigma^T X_L^T A_{\text{int}\eta}^{-1} X_L \sigma + \frac{1}{n}(B_b^2 + \eta(1 - B_b^2))\right] \\ &= \frac{1}{4n}\text{trace}(X_L^T A_{\text{int}\eta}^{-1} X_L) + \frac{1}{n}(B_b^2 + \eta(1 - B_b^2)).\end{aligned}$$

We can minimize the right hand side over $\eta \in [0, 1]$ to get the desired result. □

5.6 Proof of Theorem 7

Proof The core idea of the proof is to come up with an intuitive upper bound on the empirical Rademacher complexity of $\mathcal{F}$ using convex duality. We have already seen the use of convex duality in Proposition 1 and Theorem 6. Recall the definition of the empirical Rademacher complexity of a function class $\mathcal{F}$:

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

where $\{\sigma_i\}_{i=1}^n$ are i.i.d. Bernoulli random variables taking values in $\{\pm 1\}$ with equal probability. Now define a new vector g to be the random vector $\sum_{i=1}^{n} \sigma_i x_i$. As in the previous proofs, instead of directly working with the empirical Rademacher complexity, we will dualize the two maximization problems in the upper bound given by Eq. (4) of Lemma 1. Let $\omega(g, \mathcal{F}) = \sup_{\beta \in \mathcal{F}} g^T \beta$. That is, $\omega(g, \mathcal{F})$ is the optimal value of the first maximization problem (ignoring factor $1/n$) appearing on the right hand side of Eq. (4):

$$\begin{aligned}\max_{\beta}\ \ & g^T\beta \quad \text{s.t.} \\ & \beta^T\beta \le B_b^2 \\ & \|A_k\beta\|_2 \le a_k^T\beta + d_k \ \ \forall k = 1, ..., K. \end{aligned} \tag{13}$$

The Lagrangian of the problem can be written as (Boyd and Vandenberghe 2004):

$$\mathcal{L}(\beta, \gamma, \{z_k, \theta_k\}_{k=1}^K) = g^T\beta + \gamma(B_b^2 - \beta^T\beta) + \sum_{k=1}^{K}\Big[z_k^T A_k\beta + \theta_k \cdot (a_k^T\beta + d_k)\Big],$$

where $\beta \in \mathbb{R}^p, \gamma \in \mathbb{R}_+$ and for $k = 1, ..., K$ we have $\|z_k\|_2 \le \theta_k$. For any set of feasible values of $(\beta, \gamma, \{z_k, \theta_k\}_{k=1}^K)$, the objective of the SOCP in Eq. (13) is upper bounded by $\mathcal{L}(\beta, \gamma, \{z_k, \theta_k\}_{k=1}^K)$. Thus, $\omega(g, \mathcal{F}) \le \sup_\beta \mathcal{L}(\beta, \gamma, \{z_k, \theta_k\}_{k=1}^K)$. We will analyze this maximization problem as the first step towards a tractable bound on $\omega(g, \mathcal{F})$.

In the second step, we will minimize $\sup_\beta \mathcal{L}(\beta, \gamma, \{z_k, \theta_k\}_{k=1}^K)$ over variable γ (one of the dual variables) to get an upper bound on $\omega(g, \mathcal{F})$ in terms of $\{z_k, \theta_k\}_{k=1}^K$. These two steps are shown below:

First step: After rearranging terms and completing squares, we get the following dual objective to be minimized over dual variables γ and $\{z_k, \theta_k\}_{k=1}^K$.

$$\begin{aligned}
&\sup_{\beta\in\mathbb{R}^p} \mathcal{L}(\beta, \gamma, \{z_k, \theta_k\}_{k=1}^K) \\
&\quad= \sup_{\beta\in\mathbb{R}^p} \left[\left(g + \sum_{k=1}^K (A_k^T z_k + \theta_k a_k) \right)^T \beta + \gamma B_b^2 + \sum_{k=1}^K \theta_k d_k - \gamma \beta^T \beta \right] \\
&\quad= \sup_{\beta\in\mathbb{R}^p} \left[-\gamma \left\| \beta - \frac{g + \sum_{k=1}^K (A_k^T z_k + \theta_k a_k)}{2\gamma} \right\|_2^2 \right. \\
&\qquad \left. + \frac{\| g + \sum_{k=1}^K (A_k^T z_k + \theta_k a_k) \|_2^2}{4\gamma} + \left(\gamma B_b^2 + \sum_{k=1}^K \theta_k d_k \right) \right] \\
&\quad= \frac{\| g + \sum_{k=1}^K (A_k^T z_k + \theta_k a_k) \|_2^2}{4\gamma} + \gamma B_b^2 + \sum_{k=1}^K \theta_k d_k.
\end{aligned}$$

The second to last equality above is obtained by completing the squares (in terms of β) and the last equality is due to the fact that the optimal value is obtained when $\beta = \frac{g+\sum_{k=1}^K (A_k^T z_k + \theta_k a_k)}{2\gamma}$. The resulting term is now a function of the remaining variables (γ and $\{z_k, \theta_k\}_{k=1}^K$) and serves as an upper bound to $\omega(g, \mathcal{F})$ for any feasible values of γ and $\{z_k, \theta_k\}_{k=1}^K$.

Second step: Since $\min_{x,y} f(x, y) = \min_x (\min_y f(x, y))$ when $f(x, y)$ is convex and the feasible set is convex, we now minimize with respect to γ to get the following upper bound:

$$\begin{aligned}
&\inf_{\gamma\in\mathbb{R}_+} \sup_{\beta\in\mathbb{R}^p} \mathcal{L}(\beta, \gamma, \{z_k, \theta_k\}_{k=1}^K) \\
&\quad= B_b \left\| g + \sum_{k=1}^K (A_k^T z_k + \theta_k a_k) \right\|_2 + \sum_{k=1}^K \theta_k d_k,
\end{aligned}$$

where the above statement follows because for a problem of the form $\min_{\gamma\in\mathbb{R}_+} \frac{a}{\gamma} + b\gamma + c$ with $a > 0, b > 0$, the optimal solution is $\gamma^* = +\sqrt{\frac{a}{b}}$.

Continuing, we now optimize over the remaining variables $\{z_k, \theta_k\}_{k=1}^K$ as follows:

$$\begin{aligned}
\omega(g, \mathcal{F}) &= \sup_{\beta\in\mathcal{F}} g^T \beta \\
&\leq \inf_{\{(z_k,\theta_k): \|z_k\|_2 \leq \theta_k, k=1,..,K\}} B_b \left\| g + \sum_{k=1}^K (A_k^T z_k + \theta_k a_k) \right\|_2 + \sum_{k=1}^K \theta_k d_k. \qquad (14)
\end{aligned}$$

An upper bound on $\omega(g, \mathcal{F})$ can be obtained by finding a set of optimal or feasible values for $\{z_k, \theta_k\}_{k=1}^K$. Note that since $A_k \succ 0$, $A_k^T = A_k$ and A_k^{-1} exists. Obtaining the optimal value of the minimization in Eq. (14) is difficult analytically. Instead, we will pick a suitable feasible value for $\{z_k, \theta_k\}_{k=1}^K$. Plugging this feasible value will give us an upper bound on

$\omega(g, \mathcal{F})$. In particular, let $z_k = -\frac{1}{K} A_k^{-1} g$. Then, setting $\theta_k = \frac{1}{K}\|A_k^{-1} g\|_2$ gives us a feasible value for each $\{z_k, \theta_k\}$. Thus,

$$\begin{aligned}
\omega(g, \mathcal{F}) &\leq B_b \left\| g + \sum_{k=1}^{K} A_k^T \left(-\frac{1}{K} A_k^{-1} g \right) + \sum_{k=1}^{K} \frac{1}{K} \|A_k^{-1} g\|_2 a_k \right\|_2 + \sum_{k=1}^{K} \frac{1}{K} \|A_k^{-1} g\|_2 d_k \\
&= B_b \left\| g - g + \sum_{k=1}^{K} \frac{\|A_k^{-1} g\|_2}{K} a_k \right\|_2 + \sum_{k=1}^{K} \frac{\|A_k^{-1} g\|_2}{K} d_k \\
&= B_b \left\| \sum_{k=1}^{K} \frac{\|A_k^{-1} g\|_2}{K} a_k \right\|_2 + \sum_{k=1}^{K} \frac{\|A_k^{-1} g\|_2}{K} d_k \\
&\leq \sum_{k=1}^{K} \frac{\|A_k^{-1} g\|_2}{K} (B_b \|a_k\|_2 + d_k) \\
&\leq \|g\|_2 \sum_{k=1}^{K} \frac{B_b \|a_k\|_2 + d_k}{K \cdot \lambda_{\min}(A_k)}.
\end{aligned}$$

Dualizing the second maximization problem in Eq. (4) also gives us the same upper bound as obtained above for $\omega(g, \mathcal{F})$. That is, if $\omega'(g, \mathcal{F}) := \sup_{\beta \in \mathcal{F}} -g^T \beta$, then the same analysis as above (replacing g with $-g$) gives:

$$\omega'(g, \mathcal{F}) \leq \|g\|_2 \sum_{k=1}^{K} \frac{B_b \|a_k\|_2 + d_k}{K \cdot \lambda_{\min}(A_k)}.$$

We can now come up with the desired upper bound for the empirical Rademacher complexity using Eq. (4):

$$\begin{aligned}
\bar{\mathcal{R}}(\mathcal{F}_{|S}) &\leq \mathbb{E}\left[\max\left(\frac{1}{n}\omega(g, \mathcal{F}), \frac{1}{n}\omega'(g, \mathcal{F}) \right) \right] \\
&\leq \frac{1}{n}\mathbb{E}\left[\|g\|_2 \sum_{k=1}^{K} \frac{B_b \|a_k\|_2 + d_k}{K \cdot \lambda_{\min}(A_k)} \right] \quad \text{(since upper bounds are the same)} \\
&= \frac{1}{n}\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^{n} \sigma_i x_i \right\|_2 \right] \sum_{k=1}^{K} \frac{B_b \|a_k\|_2 + d_k}{K \cdot \lambda_{\min}(A_k)} \\
&\leq \frac{1}{n}\sqrt{\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^{n} \sigma_i x_i \right\|_2^2 \right]} \sum_{k=1}^{K} \frac{B_b \|a_k\|_2 + d_k}{K \cdot \lambda_{\min}(A_k)} \quad \text{(by Jensen's inequality)} \\
&\leq \frac{X_b}{\sqrt{n}} \sum_{k=1}^{K} \frac{B_b \|a_k\|_2 + d_k}{K \cdot \lambda_{\min}(A_k)}.
\end{aligned}$$

In the case when there are no active conic constraints, we cannot use this bound. Instead, we can recover the well known standard bound by removing the terms related to conic constraints in Eq. (14) and obtain only $\frac{X_b B_b}{\sqrt{n}}$. Combining both bounds we get,

$$\bar{\mathcal{R}}(\mathcal{F}_{|S}) \leq \frac{X_b}{\sqrt{n}} \cdot \min\left\{ B_b, \sum_{k=1}^{K} \frac{B_b \|a_k\|_2 + d_k}{K \cdot \lambda_{\min}(A_k)} \right\}.$$

□

6 Conclusion

In this paper, we have outlined how various side information about a learning problem can effectively help in generalization. We focused our attention on several types of side information, leading to linear, polygonal, quadratic and conic 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.

Appendix: Quantifying the impact of side knowledge

Here we describe an experiment[2] that we did to demonstrate the impact of side knowledge encoded as polygonal (which subsumes linear), quadratic and conic constraints. Our goal was to compare predictive accuracies of a model that used side knowledge to a baseline model that did not use side knowledge.

Algorithm setups and performance measure: We measured performance in terms of RMSE (Root Mean Squared Error) for models obtained from five setups: (1) multiple linear regression, (2) ridge regression, (3) ridge regression with polygonal constraints, (4) ridge regression with convex quadratic constraints, and (5) ridge regression with multiple conic constraints.

Dataset: The dataset for this problem was generated using a multidimensional Gaussian distribution (with a fixed covariance matrix). The number of features was set to 60. A coefficient vector was arbitrarily chosen and the response variable was computed as a linear function of the coefficient vector and the feature vector with some additional Gaussian noise. Three types of samples (feature-label pairs) were generated: (a) A test sample of size 750 was kept aside during learning. The prediction performance numbers reported in Fig. 3 were computed on this sample. (b) A "knowledge sample" of size 120 was generated in order to incorporate side knowledge as polygonal, quadratic and conic constraints. For all three types of side knowledge, the same "knowledge sample" was used, but different side knowledge was derived from it for the different algorithm setups. For polygonal (or multiple linear) constraints, a poset constraint (see Sect. 2.1) of the form $\beta^T(\tilde{x}_i - \tilde{x}_j) \leq \tilde{y}_i - \tilde{y}_j$ was constructed for each pair of points in the knowledge set and a subset were chosen for use in the convex formulation (1200 linear constraints out of a possible 7140). A quadratic constraint of the form $\|\Gamma X_U^T \beta\|_2^2 \leq c$ was constructed to impose a smoothness side knowledge (see Sect. 2.2). For this, the examples in the knowledge set were first sorted according to $\tilde{y}_i$ to be monotonic and the rows of X_U^T were reordered accordingly before being used in the constraint. The right hand side parameter c of the quadratic constraint was defined to be $\sum_{i=1}^{119}(\tilde{y}_i - \tilde{y}_{i+1})^2$ and Γ was a 119×120 matrix with $\Gamma_{i,i} = 1$ and $\Gamma_{i,i+1} = -1$ for $i = 1, .., 119$. One conic constraint for each example in the knowledge set was generated of the form $\beta^T \tilde{x}_i + r\|\beta\|_2 \leq \tilde{y}_i + r\|\beta^*\|_2$ (see Sect. 2.3). Here, the parameter r was a fixed positive real number and β^* is the true underlying coefficient vector. Knowledge of the true underlying coefficient vector is not necessary to impose such conic constraints in practice (and was used here for ease of simulation only). (c) Thirty separate training samples of size 750 were generated. Thus, each time a model was trained, it was trained on one of 30 training sets, using constraints derived from the "knowledge sample" (if it was an algorithm setup that used side knowledge) and tested on the test set.

---

[2] The source code is available at https://github.com/thejat/supervised_learning_with_side_knowledge.

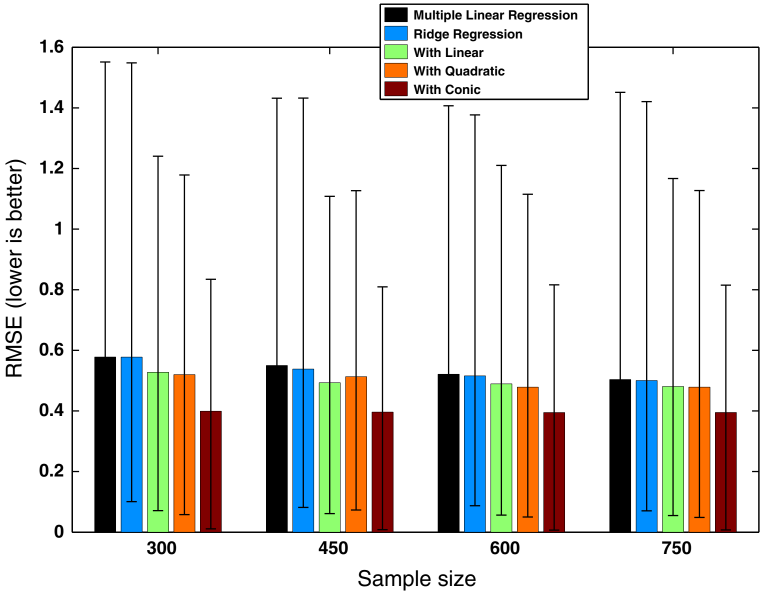

Fig. 3 Plot of predictive performance (RMSE) of models learned using different learning formulations (see the legend). Models learned using side knowledge outperform the baseline multiple linear regression and ridge regression models as evidenced by the downward shift in the 25th-75th quantile ranges (shown as whiskers on the median bar-plots). For each sample size (300, 450, 600, 750), 30 training samples were generated and used to learn 30 different models in each modeling setup (with and without the various forms of side knowledge)

Experimental Setup: For each training sample (there are 30 of them), and for each of the 5 setups, we constructed a model by solving a convex program. (For the ridge regression methods, we also performed 5-fold cross validation to choose the hyper-parameter corresponding to the ℓ_2-norm regularization term.) We then evaluated each model on the test sample and computed the RMSE. Further, to show dependence on training set size, for each training sample, we changed the data that we used from 300 examples to the full 750 examples (4 training set sizes - 300, 450, 600, 750). In summary, we learned (5 algorithm setups)*(4 training set sizes)*(30 training sets) = 600 models in this experiment, not including cross validation. Figure 3 shows the median RMSE (with 25th and 75th quantiles as whiskers) that we obtain across the 30 models.

Results: We expected a performance increase over standard multiple linear regression when we impose polygonal, quadratic and conic constraints. As seen from Fig. 3, this is indeed true. Most prominently, the distribution of RMSE error values shifts downwards when side knowledge is used. As the sample size increases, the difference in performance between a ridge regression model learned without side knowledge and those learned with side knowledge decreases as expected; the side knowledge becomes less useful when more data are available to learn from.

References

Balcan, M.F., & Blum, A. (2005). A PAC-style model for learning from labeled and unlabeled data. In Proceedings of conference on learning theory, (pp. 69–77). Springer.

Bartlett, P. L., & 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., & Mooney, R. J. (2006). Probabilistic semi-supervised clustering with constraints. In Semi-supervised learning, (pp. 71–98). Cambridge, MA. MIT Press.

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

Belkin, M., Matveeva, I., & Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. In Proceedings of conference on learning theory, (pp 624–638). Springer.

Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge, MA: Cambridge university press.

Chandrasekaran, V., Recht, B., Parrilo, P. A., & Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, *12*(6), 805–849.

Chang, M., Ratinov, Lev, & Roth, Dan (2008a). Constraints as prior knowledge. In ICML workshop on prior knowledge for text and language processing, pp. 32–39.

Chang, M-W., Ratinov, L-A., Rizzolo, N., & Roth, D. (2008b). Learning and inference with constraints. In AAAI conference on artificial intelligence, pp. 1513–1518.

Fung, G. M., Mangasarian, O. L., & Shavlik, J. W. (2002). Knowledge-based support vector machine classifiers. In Proceedings of neural information processing systems, pp. 521–528.

Gómez-Chova, L., Camps-Valls, G., Munoz-Mari, J., & Calpe, J. (2008). Semisupervised image classification with laplacian support vector machines. *Geoscience and Remote Sensing Letters, IEEE*, *5*(3), 336–340.

James, G. M., Paulson, C., & Rusmevichientong, P. (2014). The constrained lasso. working paper.

John, F. (1948). Extremum problems with inequalities as subsidiary conditions. Studies and essays presented to R. Courant on his 60th Birthday, January 8, 1948, pp. 187–204.

Kääriäinen, M. (2005). Generalization error bounds using unlabeled data. In Proceedings of conference on learning theory, (pp. 127–142). Springer.

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

Kakade, S.M., Sridharan, K., & Tewari, A. (2008). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. Proceedings of neural information processing systems, p. 22.

Andrey Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, *14*(2), 3–86.

Lanckriet, G. R. G., Ghaoui, L. E., Bhattacharyya, C., & Jordan, M. I. (2003). A robust minimax approach to classification. *The Journal of Machine Learning Research*, *3*, 555–582.

Lauer, F., & Bloch, G. (2008). Incorporating prior knowledge in support vector machines for classification: A review. *Neurocomputing*, *71*(7), 1578–1594.

Le, Quoc V., Smola, Alex J., & Gärtner, Thomas (2006). Simpler knowledge-based support vector machines. In Proceedings of the 23rd international conference on Machine learning, (pp. 521–528). ACM.

Vandenberghe, L., Boyd, S., & Lebret, H. (1998). Applications of second-order cone programming. *Linear algebra and its applications*, *284*(1), 193–228.

Lu, Z., & Leen, T. K. (2004). Semi-supervised learning with penalized probabilistic clustering. In Proceedings of neural information processing systems, pp. 849–856.

Maurer, A. (2006). The Rademacher complexity of linear transformation classes. In Proceedings of conference on learning theory, (pp. 65–78). Springer.

Nguyen, N., & Caruana, R. (2008a). Classification with partial labels. In Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, (pp. 551–559). ACM.

Nguyen, N., & Caruana, R. (2008b). Improving classification with pairwise constraints: a margin-based approach. In Machine learning and knowledge discovery in databases, (pp. 113–124). Springer.

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., & Weinshall, D. (2004). Computing Gaussian mixture models with EM using equivalence constraints. *In Proceedings of Neural Information Processing Systems*, *16*, 465–472.

Pannagadatta, K. S., Chiranjib, B., & Alexander, J. S. (2006). Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, *7*, 1283–1314.

Singh, A., Nowak, R., & Zhu, X. (2008). Unlabeled data: Now it helps, now it doesn't. In Proceedings of neural information processing systems, pp 1513–1520.

Stojnic, M. (2009). Various thresholds for l1-optimization in compressed sensing. arXiv preprint arXiv:0907.3666.

Talagrand, M. (2005). *The generic chaining*. Berlin: Springer.

Towell, G. G., Shavlik, J. W., & Noordewier, M. (1990). Refinement of approximate domain theories by knowledge-based neural networks. In Proceedings of the eighth national conference on artificial intelligence, (pp. 861–866). Boston, MA.

Tsirelson, B. S., Ibragimov, I. A., & 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, pp 20–41. Springer.

Tulabandhula, T., & Rudin, C. (2013). Machine learning with operational costs. *Journal of Machine Learning Research*, *14*, 1989–2028.

Tulabandhula, T., & Rudin, C. (2014). On combining machine learning with decision making. *Machine Learning*, *97*(1–2), 33–64.

Vapnik, V. N. (1998). *Statistical learning theory* (Vol. 2). New York: Wiley.

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.