A Statistical Learning Theory Framework for Supervised Pattern Discovery

Jonathan H. Huggins[*] Cynthia Rudin[†]

Abstract

This paper formalizes a latent variable inference problem we call *supervised pattern discovery*, the goal of which is to find sets of observations that belong to a single "pattern." We discuss two versions of the problem and prove uniform risk bounds for both. In the first version, collections of patterns can be generated in an arbitrary manner and the data consist of multiple labeled collections. In the second version, the patterns are assumed to be generated independently by identically distributed processes. These processes are allowed to take an arbitrary form, so observations within a pattern are not in general independent of each other. The bounds for the second version of the problem are stated in terms of a new complexity measure, the quasi-Rademacher complexity.

1 Introduction

The problem of *supervised pattern discovery* is that of finding sets of observations that belong together, given a set of past patterns to learn from. This problem arises naturally in domains ranging from computer vision to crime data mining. We provide a formal definition and theoretical foundation for pattern discovery. Unlike the classical problem of classification, in pattern discovery we posit an infinite number of classes ("patterns"), each containing a finite number of observations. Also, in contrast to standard classification assumptions, we do not expect the observations to be chosen i.i.d. On the contrary, the observations within a pattern may be highly correlated, whereas the latent patterns are chosen i.i.d., and our goal is to locate these patterns among the full set of observations. We briefly outline three motivating examples.

The first is a problem that is faced every day by crime analysts in police departments across the world [6, 16]. These analysts spend much of their time searching for patterns of crimes within databases of crime reports. The identification of emerging patterns of crime has been a key priority of the crime analysis profession since its inception. Analysts have knowledge of past patterns of crime, which they generalize to detect new patterns in recent crimes. Each pattern thus corresponds to the (group of) people committing the crimes. Since there are an unknown number of criminals in the world, and as patterns of crimes happen in myriad ways, we cannot assume a fixed, finite number of patterns. In addition, crimes committed by the same individual are certainly not i.i.d. On the other hand, the patterns are similar enough that sometimes the analysts can identify them. Thus, instead of the usual i.i.d. assumption of observations made generally in machine learning, we might consider each observation (crime) as being generated by one of many latent processes (the criminals), chosen i.i.d. Observations generated by the same process are considered part of a single pattern, and all of the observations are visible simultaneously. Automated methods for crime pattern detection include a neural network approach [7] and a greedy pattern-building algorithm [28]. Others [14, 5] investigate the slightly easier task of finding pairs of related crimes.

As a second example, consider the following simplified perceptual organization-style problem,[1] which involves finding geometric patterns in an image (cf. [8, 29], and also [18] for an unsupervised variant). Each observation is a line segment in $\mathbb{R}^2$. A robot or human might observe an image with more than one pattern in it: say, a star, a square, and a rhombus (see Figure 1a), which are placed according to a particular probability distribution within the space. The goal is to find the patterns, where each pattern consists of a subset of observations. In this case, a single observation can only be classified in the context of the other observations. We do not know in advance what constitutes a pattern; we have only a labeled set of other patterns to learn from. There may actually be an infinite number of pattern types. For instance, while a human might quickly recognize the two patterns in Figures 1b, there are an infinite number of other possible patterns they might also recognize in some other collection of line segments.

A final example that falls within the pattern discovery framework comes from personalized medicine [10, 11]. In personalized medicine, an individual's

[*] MIT Department of EECS and CSAIL (jhuggins@mit.edu)

[†] MIT Sloan School of Management and CSAIL (rudin@mit.edu)

[1] Ideas from perceptual organization have proven to be very useful in the field of computer vision [20].

molecular and genetic profile is used to develop a specialized treatment for that person. To accomplish this, patterns must be found within individuals' molecular/genetic profiles, the progressions of their symptoms, and the results of their treatments. These patterns are used not just to decide between one or two possible treatments. Instead, a large number of treatment regimens may be discovered, with each regimen potentially applying to only a small number of patients. For example, personalized medicine approaches have found particular success in using genome-wide gene-expression data for the treatment of cancer [21, 23, 26]. "Cancer" is a highly amorphous term. While certain cancers are caused by a few well-understood gene mutations, in many cancers there are a large number of infrequent mutations that each make a small contribution to tumorigenesis [25]. For example, breast cancer is caused by hundreds if not thousands of different mutations, with only three point mutations and perhaps ten recurrent mutations occurring in more than 10% of cases [22]. Thus, flexible pattern discovery methods are required [21, 23]. For a range of personalized medicine examples and references, see the proceedings of the recent NIPS 2010 Workshop on Predictive Models in Personalized Medicine[2].

In this paper we develop a statistical learning theory framework for two versions of the problem of supervised pattern discovery, providing a theoretical foundation for applications that are already used in practice for pattern discovery. In particular, we develop uniform risk bounds that can be used for empirical risk minimization [24]. We call the first version of the pattern discovery problem *block pattern discovery*. The block problem assumes there are collections of patterns. The collections are i.i.d. but the pattern-generating mechanisms within each collection are not necessarily independent. The second version is the *individual pattern problem*, in which the patterns are presented as a single collection and the pattern generating processes (with each generating a single pattern) are i.i.d.

To our knowledge, there are no other learning theory frameworks which, like pattern discovery, allow for an infinite number of latent patterns, each with a finite number of observations. Statistical learning theory for classification [24] supposes a finite number of possible classes each containing an infinite number of observations in the limit of infinite data, and many other supervised problems (e.g. supervised ranking) are similar. Supervised clustering [3, 2] similarly posits a known, finite number of clusters to which observations in some fixed data set belong. In the clustering model there is a "teacher" who provides feedback about the correctness of the proposed clustering and the clustering rule is assumed to come from some known concept class. Algorithms operating within the framework are concerned with finding the true rule using a polynomial number of queries to the teacher. Finally, standard clustering is an *unsupervised* method [12], whereas pattern discovery is concerned with *supervised learning*.

In addition to the theoretical work just described, there is a large body of work in the statistics, machine learning, and data mining communities on clustering (mixture modeling) with an infinite number of clusters (components). A popular approach to infinite mixture modeling is based on Bayesian non-parametric models, particularly the Dirichlet process (DP). Both unsupervised [19, 17] and semi-supervised [1] approaches have been developed. Note, however, that DP-based approaches produce clusters of infinite size in the infinite data limit, whereas we shall be interested in clusters that are of finite size even in the infinite data limit.

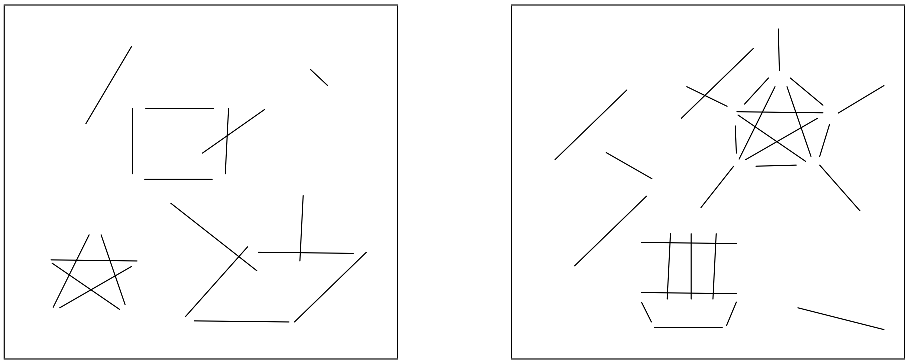
The remainder of the paper is organized as follows. In Section 2, notation is established and the block pattern discovery problem is defined. In Section 3, we give risk bounds for the block problem in terms of covering numbers. The individual pattern discovery problem is defined in Section 4 while Section 5 gives risk bounds for the individual pattern discovery problem based on an adaptation of the Rademacher complexity measure that is appropriate for pattern discovery.

2 The Block Pattern Discovery Problem

We first investigate what we call the *block pattern discovery* problem. The observations are a sequence of i.i.d. *blocks*, where a *block* is a collection of patterns. For example, a single block could be images with lines that form patterns, as in Figures 1a and 1b. In order words, each line would be an observation, a set of lines would form a geometric pattern, and the image containing the geometric patterns would constitute a block. The goal is then to find the patterns in new groups of observations, such as in new images.

Let $\mathcal{X}$ be the observation space and define $\mathcal{S}(\mathcal{X})$ to be the set of finite, non-empty subsets of $\mathcal{X}$. A pattern will consist of one or more observations from $\mathcal{X}$, so $\mathcal{S}(\mathcal{X})$ defines the collection of all possible patterns that could be observed. Also, let $\mathcal{S}^2(\mathcal{X}) := \mathcal{S}(\mathcal{S}(\mathcal{X}))$ denote all finite collections of patterns. In other words, $x \in \mathcal{X}$ is a single observation; $P \in \mathcal{S}(\mathcal{X})$ is a pattern, which consists of a finite number of observations; and $Q \in \mathcal{S}^2(\mathcal{X})$ is a finite collection of patterns (i.e. $Q = \{P_1, P_2, \dots, P_k\}, P_i \in \mathcal{S}(\mathcal{X})$). For example, in the crime pattern detection application, $x \in \mathcal{X}$ is a single crime, $P_i \in \mathcal{S}(\mathcal{X})$ are the crime patterns—crimes committed by a single person

[2] https://sites.google.com/site/personalmedmodels/



(a) Three simple patterns

(b) Two more complex patterns

Figure 1: Examples of block patterns using lines, with extra lines that do not belong to a multi-line pattern.

or group—and $Q = \{P_1, P_2, \ldots, P_k\} \in \mathcal{S}^2(\mathcal{X})$ is the collection of crime patterns. Note that each observation can belong to only one pattern.

Define an (unknown) distribution $\mathcal{D}$ over collections of patterns, so $Q \sim \mathcal{D}$ is an element of $\mathcal{S}^2(\mathcal{X})$ and can be written as $Q = \{P_1, \ldots, P_k\}$, where each $P_i \in \mathcal{S}(\mathcal{X})$ is a single pattern. Note that k is itself random. Q can be thought of as representing a labeled version of the observations — that is, indicating which observations are part of the same pattern. Let $X_Q = \bigcup_{P \in Q} P \in \mathcal{S}(\mathcal{X})$ be the set of observations associated with Q, which would correspond to an unlabeled dataset with latent patterns defined by Q.

Let $S : \mathcal{S}^2(\mathcal{X}) \to \mathcal{S}^2(\mathcal{X})$ be a *selector function*, which maps each collection of patterns $Q \in \mathcal{S}^2(\mathcal{X})$ to a subset of $\mathcal{S}(X_Q)$, the subsets of observations derived from Q. The function $S(Q)$ is used to select out which subsets of the set of observations X_Q the loss function will depend on. Since S is a function of Q, these subsets can depend on the true patterns in the observations. We will be interested in choosing a *pattern discovery function* $f : \mathcal{S}(\mathcal{X}) \times \mathcal{S}(\mathcal{X}) \to [0, 1]$. The function $f(X, U)$ outputs a score between 0 and 1, where 1 (resp. 0) indicates complete confidence that U is part (resp. not part) of a pattern from X. Let $\mathcal{F}$ be a family of pattern discovery functions. We assume throughout that for $f \in \mathcal{F}$, $f(X, U) = 0$ if $U \not\subseteq X$, since in this case it is obvious that U cannot be part of a pattern from X.

The *block loss functional* $\mathcal{L}_S$ under selector S measures the performance of the pattern discovery function $f \in \mathcal{F}$ on a collection of patterns $Q \in \mathcal{S}^2(\mathcal{X})$ and is defined to be

$$\mathcal{L}_S(f; Q) = \left[\frac{1}{Z_{Q,S}} \sum_{U \in S(Q)} \ell_{sp}^2(f, U, Q) \right]^{1/2}, \tag{2.1}$$

where $Z_{Q,S} = |S(Q)|$ is the normalization function so $\mathcal{L}_S \in [0, 1]$ and $\ell_{sp}(f, U, Q) \in [0, 1]$ is the *local loss* of f on a subset U when the true pattern collection is Q. In the following section, we will focus on the case where $\ell_{sp}(f, U, Q) = |\mathbb{I}(U \subset P \in Q) - f(X_Q, U)|$, which penalizes f for how far it is from the indicator of whether U is part of a pattern.

One particular case of interest is when the selector function is $A : Q \mapsto \mathcal{S}(X_Q)$, which selects all subsets of the data. This maximal selector can be thought of as the ideal one, in the sense that $\mathcal{L}_A$ considers the performance of f on all possible subsets. However, since there are an exponential number of subsets, evaluating $\mathcal{L}_A$ is not usually practical. In such cases, a selector that picks out subsets deemed "important" could be used. For example, the selector might choose all subsets of true patterns as positive examples and subsets of true patterns with one additional data point not from that pattern as negative examples. This particular selector function is useful for training greedy algorithms that build patterns incrementally. These kinds of greedy algorithms for finding patterns have been used both in supervised settings (e.g. that of [28]) and unsupervised settings (e.g. work on set expansion [27, 13]).

3 Risk Bounds for the Block Problem

Let the true risk for $f \in \mathcal{F}$ be

$$R(f) = \mathbb{E}_{Q \sim \mathcal{D}} \mathcal{L}_S(f; Q) \tag{3.2}$$

and the empirical risk for f given $Q_1, \ldots, Q_n \overset{i.i.d.}{\sim} \mathcal{D}$ be

$$\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_S(f; Q_i). \tag{3.3}$$

We develop bounds on the difference $L_n(f) = R(f) - \hat{R}_n(f)$ between the true and empirical risk. These bounds adapt and expand on classical learning theory results to the new pattern detection problem. Indeed, our main goal in this section is to point out that certain pattern discovery problems can be framed such that they inherit standard i.i.d. learning theory guarantees.

Because we have now packaged the block discovery problem into a learning theoretic framework, we can apply Rademacher complexity bounds, which we then relate to empirical ℓ_2 covering numbers via Dudley's entropy bound. Define the loss class to be $\mathcal{G} := \mathcal{L}_S \circ \mathcal{F} := \{g = \mathcal{L}_S(f;\cdot) \,|\, f \in \mathcal{F}\}$, and $\tilde{\mathcal{G}} = \{Q \mapsto \mathcal{L}_S(f;Q) - \mathcal{L}_S(0;Q) \,|\, f \in \mathcal{F}\}$ to be the offset loss class. Also define the empirical metric

$$d_n(g_1, g_2) = \left[n^{-1} \sum_{i=1}^{n} (g_1(Q_i) - g_2(Q_i))^2\right]^{1/2}. \tag{3.4}$$

Definition 3.1. *For metric space (T, d), the* **ϵ-covering number** *$\mathcal{N}(T, d, \epsilon)$ is defined to be the smallest integer K such that there are points $x_1, \ldots, x_K \in T$ satisfying $\bigcup_{i=1}^{K} B_\epsilon(x_i) \subseteq T$, where $B_\epsilon(x_i)$ is the open ball in T of radius ϵ centered at x_i.*

Theorem 3.1. *Let $Q_1, \ldots, Q_n \overset{i.i.d.}{\sim} \mathcal{D}$. Then for any positive integer n and any $0 < \delta < 1$, with probability $1 - \delta$ over samples of length n, every $f \in \mathcal{F}$ satisfies*

$$L_n(f) \leq \mathbb{E} \int 24 \sqrt{\frac{\log \mathcal{N}(\tilde{\mathcal{G}}, d_n, \epsilon)}{n}} \, d\epsilon + \sqrt{\frac{8 \ln(2/\delta)}{n}}. \tag{3.5}$$

Proof. We combine a Rademacher complexity bound with Dudley's entropy bound, in the Supplemental Material.

We next derive a risk bound in terms the covering number for the underlying class $\mathcal{F}$ of pattern discovery functions. In some sense this is a more natural covering number to consider than the covering number of the shifted loss class $\tilde{\mathcal{G}}$ used in Theorem 3.1, since the ultimate goal is to choose a function from $\mathcal{F}$, not $\tilde{\mathcal{G}}$. To obtain a relationship between covering numbers for $\mathcal{F}$ and $\tilde{\mathcal{G}}$, we must define a metric on $\mathcal{F}$ and relate it to the metric on $\tilde{\mathcal{G}}$. First, for $f_1, f_2 \in \mathcal{F}$, define the (squared) metric for a single collection to be (with $Z := Z_{Q,S}$)

$$\ell_Q^2(f_1, f_2) := \frac{1}{Z} \sum_{U \in S(Q)} [f_1(X_Q, U) - f_2(X_Q, U)]^2.$$

The empirical metric we are interested in is

$$\ell_n(f_1, f_2) := \left[n^{-1} \sum_{i=1}^{n} \ell_{Q_i}^2(f_1, f_2)\right]^{1/2}.$$

Theorem 3.2. *Under the same hypotheses as Theorem 3.1,*

$$L_n(f) \leq \mathbb{E} \int 24 \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \ell_n, \epsilon)}{n}} \, d\epsilon + \sqrt{\frac{8 \ln(2/\delta)}{n}}.$$

Proof. See the Supplementary Material.

Hence, we can perform empirical risk minimization in the block pattern discovery framework.

4 The Individual Pattern Discovery Problem

The block formulation of the pattern discovery problem applies to the examples from vision, crime, and medicine outlined in the introduction, but in many cases the structure of the problem may be different. Instead of working from blocks of patterns, and learning across blocks, we might wish to learn across patterns within a single block. In the crime example, each pattern collection Q_i represents an entire set of crimes, perhaps from several cities or different intervals of time. For a police department wanting to evaluate their pattern detection ability on individual patterns of crime, rather than crimes within blocks, we should try to predict when only a single pattern collection Q is available.

In the *individual pattern discovery* problem, the learner must use one collection Q as training data instead of multiple collections $Q_1, \ldots, Q_n$. The task is then to partition newly observed data $X \in \mathcal{S}(\mathcal{X})$ into patterns. We assume the finite patterns arise from stochastic processes that are chosen i.i.d. from an unknown probability distribution over processes. We allow these processes to take an arbitrary form (since, for example, a single criminal's crimes certainly are not i.i.d.). The task is to identify patterns in the observations that are not labeled as such. The individual pattern problem can thus be viewed as a kind of supervised, latent variable problem, where the pattern generating processes are the latent variables.

To formally define the individual pattern problem, let $\mathcal{P}$ be a distribution over patterns, so if $P \sim \mathcal{P}$, then $P \in \mathcal{S}(\mathcal{X})$. We are given data $P_1, \ldots, P_n \overset{i.i.d.}{\sim} \mathcal{P}$, which together form a collection of patterns $Q = \{P_1, \ldots, P_n\}$. Note that i now indexes over patterns and n denotes the number of patterns, not the number of pattern collections. Although the processes could themselves be random, since we assume all observations from each process are part of Q, all of the randomness of the processes can be absorbed into $\mathcal{P}$. It is therefore safe to equate each process with the pattern it generates.

As in the block case, we wish to choose a pattern discovery function $f : \mathcal{S}(\mathcal{X}) \to [0, 1] \in \mathcal{F}$ to minimize a loss functional, though now f does not have access to the whole set of observations X_Q, only the subset

$U \subset X_Q$ that it is making a decision on. Writing X in place of X_Q when the underlying partition is unknown, the loss function is defined to be:

$$\text{(4.6)} \qquad \mathcal{L}_\alpha(f;P,X) := \frac{\alpha}{Z_P}\sum_{U\in\mathcal{S}(P)} \ell_+(f,U) + \frac{1-\alpha}{Z_{P,X}}\sum_{U\in S_-(P,X)} \ell_-(f,U),$$

where $Z_P = |\mathcal{S}(P)|$ and $Z_{P,X} = |S_-(P,X)|$ are normalization functions. The functionals $\ell_+, \ell_- \in [0,1]$ define the losses on positive and negative examples, respectively. $S_-(P,X)$ is the *negative example selector function*, which plays an analogous role to selector function for the block loss functional. The weight factor $0 < \alpha < 1$ determines the relative importance of positive examples compared to negative examples, making the loss cost sensitive. It is necessary to weight the two sums, since otherwise in the limit as $n \to \infty$, the value of the loss could be determined solely by the negative examples. This is because as $n \to \infty$, $|X| \to \infty$ and thus $|S_-(P,X)| \to \infty$ while $|\mathcal{S}(P)|$ remains finite.

As with the general selector function S for the block loss, choosing S_- to select all negative examples would involve an exponentially large (in $|X|$) number of subsets. Therefore, instead we might define

$$\text{(4.7)} \quad S_-(P,X) = \{U \cup \{x\} \,|\, U \in \mathcal{S}(P), x \in X \setminus P\}.$$

That is, we look at how f performs on sets that are almost patterns. This choice of selector is particularly relevant for greedy algorithms, and is used by Wang et al. [28]. We will assume that S_- takes the form of equation (4.7), though our results can easily be adapted to other choices of S_- that treat the elements of X uniformly.

5 Risk Bounds for the Individual Pattern Discovery Problem

We will prove two results inspired by Bartlett and Mendelson [4]. We will first introduce a new Rademacher complexity-like quantity to use in place of the covering number term. We will also show this quantity can be well-estimated empirically. As before, we define the empirical risk

$$\text{(5.8)} \qquad \hat{R}_{\alpha,n}(f) := n^{-1}\sum_{j=1}^{n} \mathcal{L}_\alpha(f,P_j,X_Q),$$

the true risk $R_{\alpha,n}(f) := \mathbb{E}\hat{R}_{\alpha,n}(f)$, and denote the difference by $L_{\alpha,n}(f) := \hat{R}_{\alpha,n}(f) - R_{\alpha,n}(f)$. Note that unlike with the empirical risk $\hat{R}_n$ in the block problem, the terms in the sum defining $\hat{R}_{\alpha,n}(f)$ are not independent since they are all a function of X_Q. Recall that $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}_S(f;Q_i)$, where the Q_i's are all independent of each other. Let

$$\text{(5.9)} \quad \hat{\mathcal{Q}}_n(\mathcal{L}_\alpha,\mathcal{F}) := \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}} \left|n^{-1}\textstyle\sum_{i=1}^n \varepsilon_i \mathcal{L}_\alpha(f,P_i,X_Q)\right| \Big| Q\right],$$

where the ε_i are independent uniform $\{\pm 1\}$-valued random variables. Define the *quasi-Rademacher complexity* to be

$$\text{(5.10)} \qquad \mathcal{Q}_n(\mathcal{L}_\alpha,\mathcal{F}) := \mathbb{E}_Q\hat{\mathcal{Q}}_n(\mathcal{L}_\alpha,\mathcal{F}).$$

The quasi-Rademacher complexity is distinct from standard Rademacher complexity because the terms in the sum defining $\hat{\mathcal{Q}}_n(\mathcal{L}_\alpha,\mathcal{F})$ are dependent via X_Q. Hence, like Rademacher complexity, it measures the ability of the function class $\mathcal{F}$ to fit random noise, though unlike Rademacher complexity, the loss function depends on all the observations.

The two main results in this section are based on McDiarmid's inequality [15]. For independent random variables $Y_1,\ldots,Y_n$ taking values in a set V, assume that the function $F : V^n \to \mathbb{R}$ satisfies the condition that, for all $1 \le i \le n$,

$$\text{(5.11)} \quad \sup_{y_1,\ldots,y_n,y_i'\in V} |F(y_1,\ldots,y_n) - F(y_1,\ldots,y_i',\ldots,y_n)| \le c.$$

In "expectation form," McDiarmid's inequality states that for any $0<\delta<1$, with probability at least $1-\delta$,

$$\text{(5.12)} \quad F(Y_1,\ldots,Y_n) \le \mathbb{E}F(Y_1,\ldots,Y_n) + \sqrt{\frac{nc^2\log(1/\delta)}{2}}.$$

In order to apply McDiarmid's inequality, we require the following lemma.

Lemma 5.1. *Assuming $|P_i| \le B$ almost surely, then if one P_i changes, the value of $\hat{R}_{\alpha,n}(f)$ changes by at most $B_\alpha/n = (1 + 2(1-\alpha)B)/n$.*

Proof. See the Supplemental Material.

To ensure generalization, we impose a constraint on the distribution of the number of observations $|P|$ in a pattern $P \sim \mathcal{P}$. The first result assumes that $|P|$ is bounded.

Theorem 5.1. *Let $P_1,\ldots,P_n \overset{i.i.d.}{\sim} \mathcal{P}$ and assume $|P_i| \le B$ almost surely. Then for any positive integer n and any $0<\delta<1$, with probability $1-\delta$ over samples of length n, every $f \in \mathcal{F}$ satisfies*

$$\text{(5.13)} \qquad L_{\alpha,n}(f) \le 2\mathcal{Q}_n(\tilde{\mathcal{L}}_\alpha,\mathcal{F}) + \sqrt{\frac{8B_\alpha^2\ln(2/\delta)}{n}},$$

where $B_\alpha = 1 + 2(1-\alpha)B$ *and* $\tilde{\mathcal{L}}_\alpha(f;\cdot,\cdot) := \mathcal{L}_\alpha(f;\cdot,\cdot) - \mathcal{L}_\alpha(0;\cdot,\cdot)$ *is the shifted loss.*

Proof. See the Supplemental Material.

Even though all of the $\mathcal{L}_\alpha$ terms in the risk are related through X_Q, we are able to control the effect of changing one P_i on the losses $\mathcal{L}_\alpha(f;P_j,X_Q)$ when $j \neq i$. This is the key to being able to design a bound for problems as complex as supervised pattern detection. If the distribution on $|P_i|$ is arbitrary, it is not in general possible to obtain bounds such as the one above. However, we can relax the assumption that the size is bounded and instead assume geometric tails for $|P_i|$. Under this weaker condition we achieve an $O(\sqrt{\log n/n})$ convergence rate instead of $O(1/\sqrt{n})$.

Theorem 5.2. *Let* $P_1,\ldots,P_n \overset{i.i.d.}{\sim} \mathcal{P}$ *and assume that there exists a natural number* B_0*, a constant* C*, and a rate* $0 < \lambda < 1$ *such that for any* $B \geq B_0$*,* $\Pr[|P_i| > B] \leq C\lambda^B$*. Furthermore, assume that*

$$B_n := \left\lceil \frac{\log(2Cn/\delta)}{\log(1/\lambda)} \right\rceil \geq B_0. \tag{5.14}$$

Then for any positive integer n *for which equation* (5.14) *holds and all* $0 < \delta < 1$*, with probability* $1-\delta$ *over samples of length* n*, every* $f \in \mathcal{F}$ *satisfies*

$$L_{\alpha,n}(f) \leq 2\mathcal{Q}_n(\tilde{\mathcal{L}}_\alpha, \mathcal{F}) + \sqrt{\frac{8B_{n,\alpha}^2 \log(4/\delta)}{n}}, \tag{5.15}$$

where

$$B_{n,\alpha} = 1 + 2(1-\alpha)B_n. \tag{5.16}$$

Proof. If we choose some fixed $B \geq B_0$, then consider the probability that the size of all patterns is at most B,

$$\begin{aligned} \Pr[|P_i| \leq B \ \forall i = 1,\ldots,n] &= (1 - \Pr[|P_i| > B])^n \\ &\geq 1 - n\Pr[|P_i| > B] \\ &\geq 1 - nC\lambda^B, \end{aligned}$$

so the hypotheses required for Theorem 5.1 hold with probability at least $1 - nC\lambda^B$. If we set $nC\lambda^B \leq \delta/2$ and make the substitution $\delta \to \delta/2$ in the statement of Theorem 5.1, by the union bound, with probability at least $1-\delta$, every $f \in \mathcal{F}$ satisfies

$$L_{\alpha,n}(f) \leq 2\mathcal{Q}_n(\tilde{\mathcal{L}}_\alpha, \mathcal{F}) + \sqrt{\frac{8B_\alpha^2 \ln(4/\delta)}{n}}.$$

Solving $nC\lambda^B \leq \delta/2$ for B implies that the minimal choice for B is

$$B_n = \left\lceil \frac{\log(2Cn/\delta)}{\log(1/\lambda)} \right\rceil.$$

The result now follows by substituting B_n for B in the expression for B_α.

Remark 1. The risk bound given in Theorem 5.1 shows that an important parameter in determining the difficulty of the pattern discovery problem is the size of the patterns. If the patterns $P_1, P_2, \ldots$ contain a small number of elements (i.e., B is small), then pattern discovery will be easier. We note that the cost paid for large B is only linear in B. In the case of Theorem 5.2 where patterns can be arbitrarily large, tighter risk bounds are possible when most of the patterns are small.

Remark 2. The theorems proven in this section are stated in terms of the number of patterns. However, risk bounds are typically given in terms of the number of observations made. While the bounds in terms of the number of patterns are tighter, they are, essentially, asymptotically equivalent to those stated in terms of the number of observations. To see this, first consider the case where $|P|$ is almost surely bounded by B and let m be the number of observations made. Then

$$m = \sum_{i=1}^{n} |P_i| \leq nB$$

Hence, the bound in Theorem 5.1 can be rewritten in terms of m, sacrificing at most a factor of $\sqrt{B}$.

Corollary 5.1. *Under the same hypotheses as Theorem 5.1, if* m *is the number of observations taken, then for any positive integer* n *and any* $0 < \delta < 1$*, with probability* $1-\delta$ *over samples of length* n*, every* $f \in \mathcal{F}$ *satisfies*

$$L_{\alpha,n}(f) \leq 2\mathcal{Q}_n(\tilde{\mathcal{L}}_\alpha, \mathcal{F}) + \sqrt{\frac{8B \cdot B_\alpha^2 \ln(2/\delta)}{m}}. \tag{5.17}$$

An analogous result for the case when $|P_i|$ has geometric tails can be state based on the following simple fact.

Lemma 5.2. *If* $|P_i|$ *has geometric tails, then the expected number of observations in a pattern is at most*

$$B_{\lambda,C} := B_0 + \frac{C\lambda^{B_0+1}}{1-\lambda}(B_0 + 1/(1-\lambda)). \tag{5.18}$$

Proof. The proof is given in the Supplementary Material and follows by standard geometric series properties.

Since the tails of $|P_i|$ are geometric, the probability of $|P_i|$ being much greater than $B_{\lambda,C}$ is exponentially small. Thus, with high probability, the bound in Theorem 5.2, restated in terms of m (as in Corollary 5.1), is only worsened by a factor of $O(\sqrt{B_{\lambda,C}})$.

5.1 Estimating $\mathcal{Q}_n$ Like Rademacher complexity, quasi-Rademacher complexity can be empirically estimated efficiently.

THEOREM 5.3. *Assuming $|P_i| \leq B_0$ almost surely for $P_i \overset{i.i.d.}{\sim} \mathcal{P}$, then for any natural number n and any $0 < \delta < 1$, with probability $1 - \delta$ over Q and ε*

$$\left|\mathcal{Q}_n(\tilde{\mathcal{L}}_\alpha, \mathcal{F}) - \sup_{f \in \mathcal{F}} |\hat{\mathcal{Q}}_n(\tilde{\mathcal{L}}_\alpha, f)|\right| \leq \sqrt{\frac{8B_\alpha^2 \ln(2/\delta)}{n}}$$

and with probability $1 - \delta$ over Q

$$\left|\mathcal{Q}_n(\tilde{\mathcal{L}}_\alpha, \mathcal{F}) - \hat{\mathcal{Q}}_n(\tilde{\mathcal{L}}_\alpha, \mathcal{F})\right| \leq \sqrt{\frac{8B_\alpha^2 \ln(2/\delta)}{n}}$$

where

$$\hat{\mathcal{Q}}_n(\tilde{\mathcal{L}}_\alpha, f) := n^{-1} \textstyle\sum_{i=1}^n \varepsilon_i \mathcal{L}_\alpha(f, P_i, X_Q).$$

Proof. A result analogous to Lemma 5.1 can be proven for $\hat{\mathcal{Q}}_n(\tilde{\mathcal{L}}_\alpha, f)$ in place of $\hat{\mathcal{R}}_n(f)$, since they are the same up to changes in signs of the terms induced by the ε_j. Thus, the proof is essentially identical. The theorem then follows from McDiarmid's inequality applied as in the proof of Theorem 5.1.

Similar bounds can be obtained in the case that $|P_i|$ has a geometric tail (as long as equation (5.14) holds) with the right hand sides of the inequalities in the previous theorem replaced by

$$\sqrt{\frac{8B_{n,\alpha}^2 \log(4/\delta)}{n}},$$

where $B_{n,\alpha}$ is given in Theorem 5.2.

6 Algorithms and Applications for Individual Pattern Discovery

The theoretical guarantees in Section 4 lead directly to the following general algorithm for individual pattern discovery. Before running this algorithm, the user:

1. chooses a parametric class of pattern discovery functions $f_\beta : \mathcal{S}(\mathcal{X}) \to [0, 1]$, where $\beta \in \Gamma$; and
2. chooses a threshold $\theta \in [0, 1]$. If a subset of observations $\hat{P}$ scores below θ, that is $f_\beta(\hat{P}) \leq \theta$ then we would not consider $\hat{P}$ a pattern.

The algorithm is as follows:

```
ALGORITHM 6.1.
Input:
      - Data consisting of the collection of
        observations $X$
      - Training patterns $P_1, \ldots, P_n$
      - Seed $S \subset X$ of a new potential pattern
        to be discovered.

Output: New pattern $\hat{P}$.

Initialize $\hat{P} = S$.

Step 1. Train the pattern discovery algorithm on the
known patterns $Q = \{P_1, \ldots, P_n\}$ , as follows:

    $\beta^* = \min_\beta \hat{R}_{\alpha,n}(f_\beta) = \min_\beta n^{-1} \sum_{j=1}^n \mathcal{L}_\alpha(f_\beta, P_j, X_Q)$

using the definitions from Section 4 to define the loss
function and selector function. Let $f^* = f_{\beta^*}$.

Step 2. Find new pattern.

while $f^*(\hat{P}) > \theta$ do
    Compute the best observation to add to the
    set $\hat{P}$:

        $\hat{x} \in \text{argmax}_x f^*(\hat{P} \cup x)$.

    if $f^*(\hat{P} \cup \hat{x}) > \theta$ then
        the pattern has a sufficiently high score,
        and we should add $\hat{x}$ to the pattern:

            $\hat{P} \leftarrow \hat{P} \cup \hat{x}$.

Return $\hat{P}$.
```

This algorithm has the advantage of being computationally tractable, and is directly motivated by our choice of selector function. In order to select the optimal parameter β^*, we need only consider subsets of the true patterns, along with an additional observation. Also, for growing new patterns, the function f^* was specifically trained to be able to distinguish observations that belong in the pattern from those that do not belong, which is ideal for this method.

6.1 Application to Crime Pattern Detection

An algorithm that is extremely similar to the one provided above was used by Wang et al. [28] to detect crime series in the city of Cambridge, MA. In that application:

- X is a set of crimes, namely housebreaks, that happened between 1996 and 2007 in Cambridge, MA. Many details of each crime are available, including the date, time, day of the week, location, type of premise (apartment, house), location of entry (window, back door), means of entry (pried, cut screen), whether the dwelling was ransacked, etc.

- $P_1, \ldots, P_n$ is a database of known crime patterns provided by the Cambridge police department that had been curated by their Crime Analysis Unit over the decade 1996-2007. Crimes in pattern P_i were all hypothesized to have been committed by the same individual or group (they are "crime series").

- $f(P)$, $P \in \mathcal{S}(\mathcal{X})$, is a nonlinear function of the details of crimes within the pattern, called "pattern-crime similarity," parameterized by a vector λ. In particular, within function f is a linear combination of similarity measures between crimes, where λ are the linear coefficients. For instance, if j is the coefficient for location, and the value of the learned λ_j is large, it means that location is an important factor in determining whether a set of crimes is indeed part of a crime series.

- A loss function that is similar to (but slightly different than) the one provided in Section 4 was used to train the algorithm on past patterns $P_1, \ldots, P_n$ along with the rest of the crimes $\mathcal{X}$ to determine values for vector λ.

- A threshold similar to θ was used to determine when to stop growing the crime pattern. In particular, when the series becomes less cohesive after adding more crimes, the series is considered to be complete.

The algorithm of Wang et al. [28] has been successful in being able to detect patterns of crime in Cambridge, and in a blind test with Cambridge crime analysts, it has been able to locate 9 crimes that belong in patterns that were not previously identified as such, and it was able to exclude 8 crimes that analysts previously thought were part of a pattern (they now agree that these crimes are not part of a pattern). In one case, the exclusion of a crime from a pattern helped to narrow the suspect description down to one possible race and gender (white male).

6.2 Application to Set Completion and "Growing a List" in Information Retrieval Another algorithm similar to Algorithm 6.1 was used for the problem of set completion in information retrieval. A "set completion engine" is a next generation search engine. It takes a few seed examples, of almost anything, and simply aims to produce more of them. For instance, a search starting with seed "Boston Harborfest" and "South Boston Street Festival" should yield a list of more large annual events in Boston. The algorithm of Letham et al. [13] for growing a list of items from a seed uses an algorithm similar to Algorithm 6.1 in that at each iteration, a new item is added to the set. Here:

- X is a set of all terms and phrases found on the internet.

- $P_1, \ldots, P_n$ is a set of gold standard completed sets, such as the "List of ..." articles on Wikipedia used for experiments of Letham et al. [13].

- $f(P)$, $P \in \mathcal{S}(\mathcal{X})$, is a linear combination of similarities between terms, with coefficients chosen for the Bayesian Sets algorithm of Ghahramani and Heller [9]. This algorithm is unsupervised, meaning that the training step in Algorithm 6.1 is replaced with a Bayesian prior. It would not be difficult to design a supervised algorithm that learns the prior hyperparameters of Bayesian Sets, rather than having the user choose them. In the experiments of Letham et al. [13], the parameters were chosen using a heuristic. The terms that are combined are indicator variables of the internet domains where items can be found.

In the case of growing a list, Letham et al. [13] showed that as long as the feature space and score f are constructed correctly, the results coming from this algorithm are accurate enough to be used in practice, and are substantially more accurate than other methods currently in use for set completion, including Boo!Wa![3] and Google Sets.[4]

The present work thus provides theoretical foundations for the methodologies used by Wang et al. [28] and Letham et al. [13] that we discussed in the Sections 6.1 and 6.2.

Acknowledgements

Thanks to Dylan Kotliar and Yakir Reshef for helpful discussions regarding personalized medicine and cancer genomics and to Peter Krafft for helpful comments. Funding provided by MIT Lincoln Laboratory and NSF-CAREER IIS-1053407. Supplementary material is available at the first author's personal website.

References

[1] F Akova, M Dundar, Y Qi, and B Rajwa. Self-Adjusting Models for Semi-supervised Learning in Partially Observed Settings. In *ICDM*, pages 21–30. IEEE, 2012.

[3] www.boowa.com

[4] Google Sets is available through Google Spreadsheet.

[2] P Awasthi and R B Zadeh. Supervised clustering. In *NIPS*, pages 91–99, 2010.

[3] M.-F. Balcan. *New Theoretical Frameworks for Machine Learning*. PhD thesis, Carnegie Mellon University, 2008.

[4] P L Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *The Journal of Machine Learning Research*, 3, 2002.

[5] Donald E. Brown and Stephen Hagen. Data association methods with applications to law enforcement. *Decision Support Systems*, 34(4):369–378, 2003.

[6] H Chen, W Chung, J J Xu, G Wang, Y Qin, and M Chau. Crime data mining: a general framework and some examples. *Computer*, 37(4):50–56, 2004.

[7] K Dahbur and T Muscarello. Classification system for serial criminal patterns. *Artificial Intelligence and Law*, 11:251–269, 2003.

[8] V Ferrari, T Tuytelaars, and L Van Gool. Object detection by contour segment networks. In *ECCV*, pages 14–28. Springer, 2006.

[9] Zoubin Ghahramani and Katherine Heller. Bayesian sets. In *NIPS*, 2005.

[10] G S Ginsburg and J J McCarthy. Personalized medicine: revolutionizing drug discovery and patient care. *TRENDS in Biotechnology*, 19(12):491–496, 2001.

[11] M A Hamburg and F S Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning*. Springer, 2008.

[13] B Letham, C Rudin, and K Heller. Growing a list. *Data Mining and Knowledge Discovery*, 2013.

[14] Song Lin and Donald E. Brown. An outlier-based data association method for linking criminal incidents. In *Proceedings of the Third SIAM International Conference on Data Mining*, 2003.

[15] C McDiarmid. Concentration. Technical report, University of Oxford, 1998.

[16] Shyam Varan Nath. Crime pattern detection using data mining. In *Web Intelligence and Intelligent Agent Technology Workshops*, pages 41–44, 2006.

[17] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[18] N Payet and S Todorovic. From a set of shapes to object discovery. *ECCV*, pages 57–70, 2010.

[19] C. E. Rasmussen. The infinite Gaussian mixture model. In *NIPS*, pages 554–560, 1999.

[20] S Sarkar. An introduction to perceptual organization. In *Integration of Knowledge Intensive Multi-Agent Systems, 2003. International Conference on*, 2003.

[21] T Sørlie, C M Perou, R Tibshirani, and ... Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98 (19):10869–10874, 2001.

[22] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[23] L J van't Veer, H Dai, M J Van De Vijver, Y D He, and ... Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2002.

[24] V Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[25] B Vogelstein, N Papadopoulos, V E Velculescu, S Zhou, L A Diaz, and K W Kinzler. Cancer Genome Landscapes. *Science*, 339(6127):1546–1558, March 2013.

[26] L Wang, M S Lawrence, Y Wan, and ... SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 365 (26):2497–2506, December 2011.

[27] R. C. Wang and W. W. Cohen. Iterative set expansion of named entities using the web. In *Proceedings of ICDM*, 2008.

[28] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Detecting patterns of crime with series finder. In *ECML-PKDD*, 2013.

[29] Q Zhu, L Wang, Y Wu, and J Shi. Contour context selection for object detection: A set-to-set contour matching approach. In *ECCV*, pages 774–787. Springer, 2008.

Supplemental Material for *A Statistical Learning Theory Framework for Supervised Pattern Discovery*

Jonathan H. Huggins[*] Cynthia Rudin[†]

A Proof of Theorem 3.1

Recall the statement of Theorem 3.1:

Theorem. *Let $Q_1, \dots, Q_n \overset{i.i.d.}{\sim} \mathcal{D}$. Then for any positive integer n and any $0 < \delta < 1$, with probability $1 - \delta$ over samples of length n, every $f \in \mathcal{F}$ satisfies*

(1)

$$L_n(f) \le \mathbb{E} \int 24 \sqrt{\frac{\log \mathcal{N}(\tilde{\mathcal{G}}, d_n, \epsilon)}{n}}\, d\epsilon + \sqrt{\frac{8 \ln(2/\delta)}{n}}.$$

Proof. Let

$$\hat{\mathcal{R}}_n(\mathcal{F}) := \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| n^{-1} \textstyle\sum_{i=1}^n \varepsilon_i f(X_i) \right| \middle| \{X_i\} \right],$$

where the ε_i are independent uniform $\{\pm 1\}$-valued random variables. Define the *Rademacher complexity* to be $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}\hat{\mathcal{R}}_n(\mathcal{F})$. We will need the following Rademacher complexity-based risk bound.

Theorem A.1 ([1]). *Let $Q_1, \dots, Q_n \overset{i.i.d.}{\sim} \mathcal{D}$. Then for any positive integer n and any $0 < \delta < 1$, with probability $1 - \delta$ over samples of length n, every $f \in \mathcal{F}$ satisfies*

$$L_n(f) \le 2\mathcal{R}_n(\tilde{\mathcal{G}}) + \sqrt{\frac{8 \ln(2/\delta)}{n}}.$$

Recall that $\mathcal{G}$ is the loss class and $\tilde{\mathcal{G}}$ the offset loss class. We can relate the Rademacher complexity term in the preceding theorem to the empirical covering number of $\mathcal{F}$. This can be done via Dudley's entropy bound, which we state just for case of the Rademacher process

$$\hat{\mathcal{R}}_n(g) = n^{-1/2} \sum_{i=1}^n \varepsilon_i g(Q_i).$$

Theorem A.2 (Dudley's entropy bound; [2]). *For the Rademacher process $\hat{\mathcal{R}}_n$ defined above,*

$$\mathbb{E} \sup_{g \in \mathcal{G}} \hat{\mathcal{R}}_n(g) \le \int 12 \sqrt{\log \mathcal{N}(\mathcal{F}, d_n, \epsilon)}\, d\epsilon.$$

Since $\mathcal{R}_n(\mathcal{G}) = \mathbb{E} \sup_{g \in \mathcal{G}} n^{-1/2} \hat{\mathcal{R}}_n(g)$, combining the two theorems gives result. □

B Proof of Theorem 3.2

Recall the statement of Theorem 3.2:

Theorem. *Under the same hypotheses as Theorem 3.1,*

$$L_n(f) \le \mathbb{E} \int 24 \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \ell_n, \epsilon)}{n}}\, d\epsilon + \sqrt{\frac{8 \ln(2/\delta)}{n}}.$$

Proof. Let $I_Q(X_Q, U) := \mathbb{I}(U \subset P \in Q)$. Then for $g \in \tilde{\mathcal{G}}$ we have

$g(Q) = \mathcal{L}_S(f; Q) - \mathcal{L}_S(0; Q) = \ell_Q(f, I_Q) - \ell_Q(0, I_Q)$, so

$$\begin{aligned} d_n(g_1, g_2) &= \left[n^{-1} \sum_{i=1}^n (g_1(Q_i) - g_2(Q_i))^2 \right]^{1/2} \\ &= \left[n^{-1} \sum_{i=1}^n (\ell_{Q_i}(f_1, I_{Q_i}) - \ell_{Q_i}(f_2, I_{Q_i}))^2 \right]^{1/2} \\ &\le \left[n^{-1} \sum_{i=1}^n \ell_{Q_i}^2(f_1, f_2) \right]^{1/2} = \ell_n(f_1, f_2). \end{aligned}$$

This inequality implies that if the ϵ-ball centered at f (with metric ℓ_n) contains f' then the ϵ-ball centered at g (with metric d_n) contains g'. Hence $\mathcal{N}(\mathcal{G}, d_n, \epsilon) \le \mathcal{N}(\mathcal{F}, \ell_n, \epsilon)$, which together with Theorem 3.1 gives the result. □

[*] MIT Department of EECS and CSAIL (jhuggins@mit.edu)

[†] MIT Sloan School of Management and CSAIL (rudin@mit.edu)

C Proof of Lemma 5.1

Recall the statement of Lemma 5.1:

Lemma. *Assuming $|P_i| \le B$ almost surely, then if one P_i changes, the value of $\hat{R}_{\alpha,n}(f)$ changes by at most $B_\alpha/n = (1 + 2(1-\alpha)B)/n$.*

Proof. Let $\Delta\mathcal{L}_\alpha^{j,i}$ denote the maximum possible change in $\mathcal{L}_\alpha(f, P_j, X_Q)$ due to a change in P_i:

$$\Delta\mathcal{L}_\alpha^{j,i} = \sup_{P_i'} |\mathcal{L}_\alpha(f, P_j, X_Q) - \mathcal{L}_\alpha(f, P_j, X_{Q'})|,$$

where $Q' = \{P_1, P_2, \ldots, P_{i-1}, P_i', P_{i+1}, \ldots, P_n\}$. Recall that $\mathcal{L}_\alpha(f; P_j, X_Q)$ consists of a sum of losses over $\mathcal{S}(P_j)$ weighted by α/Z_{P_j} and a sum of losses over $S_-(P_j, X_Q)$ weighted by $(1-\alpha)/Z_{P_j,X_Q}$.

For $j \ne i$, the sum

$$\frac{\alpha}{Z_{P_j}} \sum_{U \in \mathcal{S}(P_j)} \ell_+(f, U)$$

from $\mathcal{L}_\alpha(f, P_j, X_Q)$ remains constant when P_i changes since it does not depend on X_Q. The second normalizing constant is

$$Z_{P_j,X_Q} = (|X_Q| - |P_j|)(2^{|P_j|} - 1),$$

which is the number of nontrivial subsets of P_j combined with one element that is not from P_j.

Write

$$\begin{aligned} Z &= Z_{P_j,X_Q} \\ Z' &= Z_{P_j,X_{Q'}} \\ Y &= \sum_{U \in S_-(P_j,X_Q)} \ell_-(f,U) \\ \Delta Y &= Y - \sum_{U \in S_-(P_j,X_{Q'})} \ell_-(f,U) \end{aligned}$$

so we have

$$\begin{aligned} \left| \frac{Y}{Z} - \frac{Y - \Delta Y}{Z'} \right| &= \frac{|Z'Y - Z(Y - \Delta Y)|}{ZZ'} \\ &\le \frac{Y|Z' - Z|}{ZZ'} + \frac{|\Delta Y|}{Z'} \\ &\le \frac{|Z' - Z| + |\Delta Y|}{Z'}, \end{aligned}$$

where we used that $Y/Z \le 1$. At most $|P_i|(2^{|P_j|} - 1)$ terms in the sum

$$\sum_{U \in S_-(P_j,X_Q)} \ell_-(f,U)$$

can change value when P_i changes. This is the number of subsets of P_j combined with one element of P_i. Since $\ell_- \in [0,1]$ and $|P_i| \le B$, we therefore have

$$\Delta Y \le |P_i|(2^{|P_j|} - 1) \le B(2^{|P_j|} - 1).$$

Also,

$$|Z - Z'| = (|P_i| - |P_i'|)(2^{|P_j|} - 1) \le B(2^{|P_j|} - 1)$$

and

$$Z' = (|X_{Q'}| - |P_j|)(2^{|P_j|} - 1).$$

Combining these we get (for $j \ne i$)

$$\begin{aligned} \Delta\mathcal{L}_\alpha^{j,i} &\le (1-\alpha) \left| \frac{Y}{Z} - \frac{Y - \Delta Y}{Z'} \right| \\ &\le \frac{2(1-\alpha)B(2^{|P_j|} - 1)}{(|X_{Q'}| - |P_j|)(2^{|P_j|} - 1)} \\ &\le \frac{2(1-\alpha)B}{n-1}, \end{aligned}$$

where we have used the fact that $|X_{Q'}| - |P_j| \ge n - 1$, since there is at least one element in each pattern. We also have the trivial bound that $\Delta\mathcal{L}_\alpha^{i,i} \le 1$ because $\mathcal{L}_\alpha \in [0,1]$. Letting $\hat{R}'_{\alpha,n}(f)$ denote $\hat{R}_{\alpha,n}(f)$ when P_i is replaced by P_i', we have

$$\begin{aligned} &\sup_{P_i'} |\hat{R}_{\alpha,n}(f) - \hat{R}'_{\alpha,n}(f)| \le \\ &n^{-1} \sum_j \Delta\mathcal{L}_\alpha^{j,i} \le n^{-1} \left(1 + \sum_{j \ne i} \Delta\mathcal{L}_\alpha^{j,i} \right) \\ &\le n^{-1} \left(1 + (n-1) \frac{2(1-\alpha)B)}{n-1} \right) \\ &= \frac{1 + 2(1-\alpha)B}{n} = \frac{B_\alpha}{n}. \end{aligned}$$

□

D Proof of Theorem 5.1

Recall the statement of Theorem 5.1:

Theorem. *Let $P_1, \ldots, P_n \overset{i.i.d.}{\sim} \mathcal{P}$ and assume $|P_i| \leq B$ almost surely. Then for any positive integer n and any $0 < \delta < 1$, with probability $1 - \delta$ over samples of length n, every $f \in \mathcal{F}$ satisfies*

$$L_{\alpha,n}(f) \leq 2\mathcal{Q}_n(\tilde{\mathcal{L}}_\alpha, \mathcal{F}) + \sqrt{\frac{8B_\alpha^2 \ln(2/\delta)}{n}}, \tag{2}$$

where $B_\alpha = 1 + 2(1 - \alpha)B$ and $\tilde{\mathcal{L}}_\alpha(f; \cdot, \cdot) := \mathcal{L}_\alpha(f; \cdot, \cdot) - \mathcal{L}_\alpha(0; \cdot, \cdot)$ is the shifted loss.

Proof. We have

$$\begin{aligned} R_{\alpha,n}(f) &\leq \hat{R}_{\alpha,n}(f) + \sup_{f' \in \mathcal{F}} \big(R_{\alpha,n}(f') - \hat{R}_{\alpha,n}(f')\big) \\ &= \hat{R}_{\alpha,n}(f) + R_{\alpha,n}(0) - \hat{R}_{\alpha,n}(0) \\ &\quad + \sup_{f' \in \mathcal{F}} \big\{R_{\alpha,n}(f') - \hat{R}_{\alpha,n}(f') \\ &\qquad\qquad R_{\alpha,n}(0) - \hat{R}_{\alpha,n}(0)\big\} \\ &= \hat{R}_{\alpha,n}(f) + R_{\alpha,n}(0) - \hat{R}_{\alpha,n}(0) \\ &\quad + \sup_{f' \in \mathcal{F}} \big\{\mathbb{E}\tilde{R}_{\alpha,n}(f') - \tilde{R}_{\alpha,n}(f')\big\}, \end{aligned} \tag{3}$$

where $\tilde{R}_{\alpha,n}(f') := n^{-1} \sum_{j=1}^n \tilde{\mathcal{L}}_\alpha(f', P_j, X_Q)$.

We now wish to bound the final two terms of (3). First consider the term $\sup_{f' \in \mathcal{F}} \big(\mathbb{E}\tilde{R}_{\alpha,n}(f') - \tilde{R}_{\alpha,n}(f')\big)$. Note that $\mathbb{E}\tilde{R}_{\alpha,n}(f')$ is a constant. On the other hand, if one P_i changes, $\tilde{R}_{\alpha,n}(f')$ can change by at most $2B_\alpha/n$ since $\tilde{R}_{\alpha,n}(f') = \hat{R}_{\alpha,n}(f) - \hat{R}_{\alpha,n}(0)$ and each of these $\hat{R}_{\alpha,n}(\cdot)$ terms can change by at most B_α/n by Lemma 5.1. Now applying McDiarmid's inequality with $F = \sup_{f' \in \mathcal{F}} \big(\mathbb{E}\tilde{R}_{\alpha,n}(f') - \tilde{R}_{\alpha,n}(f')\big)$ and $c = 2B_\alpha/n$, we have that with probability at least $1 - \delta/2$,

$$\begin{aligned} &\sup_{f' \in \mathcal{F}} \big(\mathbb{E}\tilde{R}_{\alpha,n}(f') - \tilde{R}_{\alpha,n}(f')\big) \\ &\leq \mathbb{E} \sup_{f' \in \mathcal{F}} \big(\mathbb{E}\tilde{R}_{\alpha,n}(f') - \tilde{R}_{\alpha,n}(f')\big) + \sqrt{2B_\alpha^2 \ln(2/\delta)/n}. \end{aligned} \tag{4}$$

An essentially identical argument applies to bounding $R_{\alpha,n}(0) - \hat{R}_{\alpha,n}(0)$ by noting that $R_{\alpha,n}(0)$ is a constant and that $\mathbb{E}[R_{\alpha,n}(0) - \hat{R}_{\alpha,n}(0)] = 0$, so with probability at least $1 - \delta/2$,

$$R_{\alpha,n}(0) - \hat{R}_{\alpha,n}(0) \leq \sqrt{2B_\alpha^2 \ln(2/\delta)/n}. \tag{5}$$

Combining the previous two bounds with (3) gives, with probability at least $1 - \delta$,

$$\begin{aligned} R_{\alpha,n}(f) \leq \hat{R}_{\alpha,n}(f) &+ \sqrt{\frac{8B_\alpha^2 \ln(2/\delta)}{n}} \\ &+ \mathbb{E} \sup_{f' \in \mathcal{F}} \big(\mathbb{E}\tilde{R}_{\alpha,n}(f') - \tilde{R}_{\alpha,n}(f')\big). \end{aligned} \tag{6}$$

To complete the proof, let $P_1', \ldots, P_n' \overset{i.i.d.}{\sim} \mathcal{P}$ and let $Q' = \{P_1', \ldots, P_n'\}$. Now, writing what the expectations are with respect to for clarity, we have

$$\begin{aligned} &\mathbb{E}_Q \sup_{f' \in \mathcal{F}} \big(\mathbb{E}_{Q'}\tilde{R}_{\alpha,n}(f') - \tilde{R}_{\alpha,n}(f')\big) \\ &= \mathbb{E}_Q \sup_{f' \in \mathcal{F}} \mathbb{E}_{Q'} \left[n^{-1} \sum_{j=1}^n \tilde{\mathcal{L}}_\alpha(f', P_j', X_{Q'}) - \tilde{R}_{\alpha,n}(f') \middle| Q \right] \\ &\leq \mathbb{E}_{Q,Q'} \sup_{f' \in \mathcal{F}} \left[n^{-1} \sum_{j=1}^n \tilde{\mathcal{L}}_\alpha(f', P_j', X_{Q'}) - \tilde{R}_{\alpha,n}(f') \right] \\ &= \mathbb{E}_{Q,Q',\varepsilon} \sup_{f' \in \mathcal{F}} \left[n^{-1} \sum_{j=1}^n \varepsilon_i \big(\tilde{\mathcal{L}}_\alpha(f', P_j', X_{Q'}) - \tilde{\mathcal{L}}_\alpha(f', P_j, X_Q)\big) \right] \\ &\leq 2\mathbb{E}_{Q,\varepsilon} \sup_{f' \in \mathcal{F}} \left[n^{-1} \sum_{j=1}^n \varepsilon_i \tilde{\mathcal{L}}_\alpha(f', P_j, X_Q) \right] \\ &\leq 2\mathcal{Q}_n(\tilde{\mathcal{L}}_\alpha, \mathcal{F}). \end{aligned}$$

The first line follows from the definition of $\tilde{R}_{\alpha,n}(f')$. The first inequality follows from Jensen's inequality applied to sup. The second equality follows by symmetry. The second inequality follows since symmetry permits the difference in each pair of $\tilde{\mathcal{L}}_\alpha$ terms can be bounded by twice the (worst) of one term. The last inequality relies on Jensen's inequality applied to $|\cdot|$ and the fact that $|\sup \cdot| \leq \sup |\cdot|$. For more details about this type of proof technique, see [1]. □

E Proof of Lemma 5.2

Recall the statement of Lemma 5.2:

Lemma. *If $|P_i|$ has geometric tails, then the expected number of observations in a pattern is at most*

$$(7)\qquad B_{\lambda,C} := B_0 + \frac{C\lambda^{B_0+1}}{1-\lambda}(B_0 + 1/(1-\lambda)).$$

Proof. Since $\Pr[|P| \geq B] \leq C\lambda^B$ for $B \geq B_0$, $P_B := \Pr[|P| = B] \leq C\lambda^B$ for $B \geq B_0$. Note that

$$\begin{aligned} \lambda\frac{\partial}{\partial\lambda}\sum_{B=B_0}^{\infty}\lambda^B &= \lambda\frac{\partial}{\partial\lambda}\frac{\lambda^{B_0}}{1-\lambda} \\ \sum_{B=B_0}^{\infty}B\lambda^B &= \lambda\frac{B_0\lambda^{B_0-1}(1-\lambda)+\lambda^{B_0}}{(1-\lambda)^2} \\ &= \lambda^{B_0}\frac{B_0(1-\lambda)+\lambda}{(1-\lambda)^2}. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}[|P|] &\leq B_0 + \sum_{B=B_0+1}^{\infty} BP_B \\ &\leq B_0(1 - C\lambda^{B_0}) + C\sum_{B=B_0}^{\infty}B\lambda^B \\ &= B_0(1 - C\lambda^{B_0}) + C\lambda^{B_0}\frac{B_0(1-\lambda)+\lambda}{(1-\lambda)^2} \\ &= B_0 + \frac{C\lambda^{B_0+1}}{1-\lambda}(B_0 + 1/(1-\lambda)). \end{aligned}$$

Note the change in the start of the summation between the first and second line. □

References

[1] P L Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *The Journal of Machine Learning Research*, 3, 2002.

[2] S. Mendelson. Improving the sample complexity using global data. *Information Theory, IEEE Transactions on*, 48(7):1977–1991, 2002.