

---

# Learning with feature feedback: from theory to practice

---

Stefanos Poulis  
UC San Diego  
spoulis@eng.ucsd.edu

Sanjoy Dasgupta  
UC San Diego  
dasgupta@cs.ucsd.edu

## Abstract

In supervised learning, a human annotator only needs to assign each data point (document, image, etc.) its correct label. But in many situations, the human can also provide richer feedback at essentially no extra cost. In this paper, we examine a particular type of *feature feedback* that has been used, with some success, in information retrieval and in computer vision. We formalize two models of feature feedback, give learning algorithms for them, and quantify their usefulness in the learning process. Our experiments also show the efficacy of these methods.

## 1 INTRODUCTION

In supervised learning, obtaining a labeled training data set can be costly: a human labeler needs to scrutinize each data point and determine its label. One approach to reducing this expense is *active learning*: the learner intelligently and adaptively decides which points should be labeled. There are several active learning methods that work well in practice and enjoy theoretical guarantees [Settles, 2012, Dasgupta, 2011]. Here we consider a strategy complementary to active learning: can the human, while examining the data point, provide not just the label but also the identity of one or more relevant features?

Consider, for example, a document classification problem in which a labeler assigns each document  $x$  to a category  $y$  (“sports”, “politics”, and so on). While making this determination, the labeler might also be able to highlight a few words that are highly indicative of the label (e.g. “Congress”, “Obama”, “filibuster”). Figure 1a illustrates feature feedback. This additional informa-

tion can often be provided with minimal extra effort and might be useful for learning a classifier. Early work in information retrieval that advocates this auxiliary feedback is that of Croft and Das [1990]. Since then, there have been several experimental studies of different methods for exploiting this feedback [Raghavan et al., 2005, Dayanik et al., 2006, Druck et al., 2008, Raghavan and Allan, 2007, Settles, 2011].

Alternatively, consider a computer vision system that is learning to recognize different animals. Whenever it makes a mistake—classifies a “zebra” as a “horse”, say—a human labeler corrects it. While doing this, the labeler can also, at no extra cost, highlight a part of the image (the stripes, for instance) that distinguishes the two animals. Recent work on recognizing different species of birds, for instance, has used this effectively [Branson et al., 2010].

This kind of feedback is not trivial to model. For one thing, it is potentially quite ambiguous. Let’s return to the example of a document about “politics” in which the labeler highlights the word “filibuster”. This word is, indeed, predictive of the label, but it is also so specific that it might not apply to very many documents. Should “filibuster” be treated as a proxy for a whole collection of words that co-occur with it, or possibly a proxy for an entire *topic*? This seems reasonable, but what is the right level of granularity for the topic, or the cluster of co-occurring words?

Similarly, in the computer vision example, suppose a labeler decides that a bird is a particular type of robin and provides additional feedback by clicking on its breast (whose color, for instance, might be a deciding factor). The learner may have some higher-level representation of the image, for instance a hierarchical parts decomposition, in which case it will in general be unclear which of these features the user is referring to: several features, at different scales, might be candidates.

In both the text classification and vision examples, we see that there is the raw input  $x$  (document, image), as well as an intermediate representation  $z$  (clusters

of words, hierarchical parts decomposition) that the labeler can not access directly. After deciding on the label  $y$ , the labeler indicates one or more coordinates in  $x$ ; these indirectly and noisily reference a subset of features in  $z$ , of which some might be relevant to  $y$  and some not. Under this scenario there is some *vagueness* in the labeler’s intent: he/she directly acts on coordinates in  $x$  whose selection triggers a subset of coordinates in  $z$ . Figure 1b illustrates vague feature feedback for the example of document classification.

**Contributions.** In contrast to the richness of theoretical results for label-based active learning, there is little theoretical work on feature feedback. In this paper we formalize two models of feature feedback and give learning algorithms for them, along with theoretical guarantees. We also provide experimental corroboration of the efficacy of these methods.

The first model we study is a probabilistic generalization of disjunctions. For concreteness, we define this model specifically in the document-topic setting, but it applies more generally to the  $x$ - $z$ - $y$  situation described above: the label  $y$  of each document  $x$  is assumed to be probabilistically generated from the unnamed intermediate-level features  $z$ . We call this the *probabilistic disjunction model* (PDM). If we only had documents and labels, we could try to find a maximum-likelihood fit for the generative model, but we show that this is an NP-hard problem. On the other hand, feature feedback makes learning tractable. We give an efficient algorithm that exploits this feedback to learn a PDM. Under simple assumptions, we are able to guarantee the correctness of this algorithm, as well as quantify its label complexity.

The PDM model is not as expressive as linear separators, which are commonly used in document classification. To address this, the second situation we study is learning linear separators from feature feedback. We suggest a straightforward approach to incorporating information that a particular feature is relevant: reducing the degree of regularization on that feature. This is algorithmically simple and we show that it leads to better generalization bounds.

The regularization approach to feature feedback has the drawback of not directly modeling vagueness in the labeler’s intent. We incorporate this in a bootstrapped PDM algorithm in which a PDM is first fit to data, using a small amount of feature feedback, and is then used to label whichever documents it is confident about. This augmented training set is then used to train a linear separator (or any other model of interest).

The rest of the paper is organized as follows. In section 2 we review previous work in learning with feature

feedback. In section 3 we lay out the problem of learning probabilistic disjunctions. We give a hardness result and in turn, a simple and efficient algorithm for learning a PDM. Then, in section 4 we study linear separators. We first provide an improved generalization bound for feature feedback, and then derive a practical algorithm for learning a linear support vector machine with feature feedback (SVM-FF). To cope with the limitations of the PDM and SVM-FF we propose the bootstrapped PDM. In section 6 we present a variety of simulation experiments comparing these methods (PDM, SVM-FF and bootstrapped PDM) on several benchmark text categorization data sets. We then conduct a user study to assess feature feedback in a situation with human annotators and conclude.

## 2 RELATED WORK

There is a lot of work on incorporating domain knowledge into learning, for instance by using this knowledge to construct a preliminary classifier or to set Bayesian hyperparameters (Schapire et al. [2002], Wu and Srihari [2004], and Dayanik et al. [2006]).

For feature feedback more specifically, the feedback model closest in spirit to ours is probably that of Druck et al. [2007], whose *generalized expectation criteria* framework incorporates user-supplied feature-label relationships into the objective function for learning. Another line of work develops the idea of *annotator rationales* (Zaidan et al. [2007], Zaidan and Eisner [2008], Donahue and Grauman [2011]), in which the labeler highlights regions of the document that serve as explanations of the label; these are then used to generate *contrast examples* (same document, but with these regions removed) and the learning procedure asks for each document to be distinguished from its contrasting version. This framework involves denser annotation than we have in mind. A related form of “contrast example” is considered by Sun and DeJong [2005], who incorporate this into an SVM framework and provide generalization bounds—though these are weaker and less general than our bounds, which have less requirements on the feedback and apply to any linear model. Later work by Small et al. [2011] developed the *constrained weight-space* SVM framework by allowing annotators to provide ranked features. One further research thread includes work developed in Melville et al. [2004, 2005], Raghavan et al. [2006], Sindhwani et al. [2009], where active learning is used to incorporate feature feedback into learning. The framework there is to identify the *most informative features* to be shown to the human, when asked to label an example.

In the above works feature feedback is explicit: information about particular (feature, label) associations does

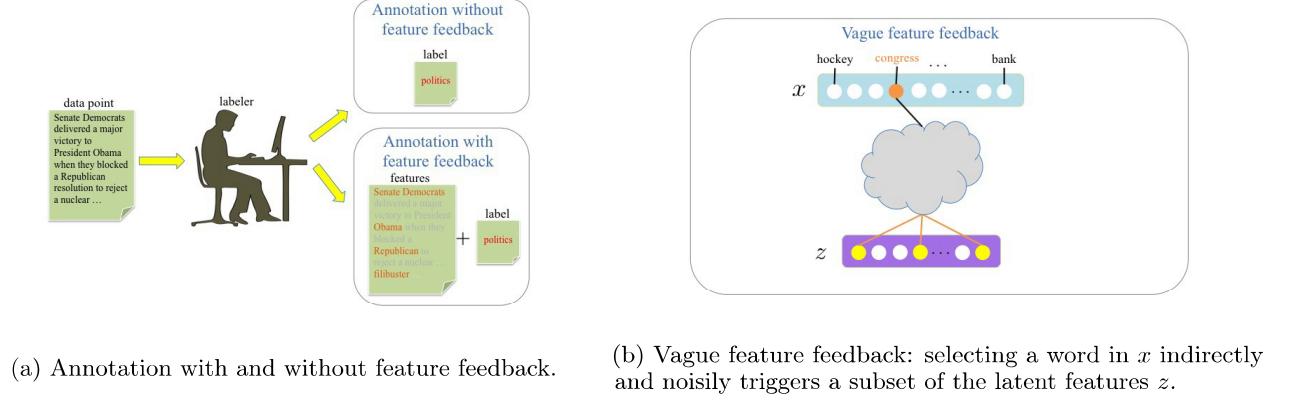


Figure 1: Models of Feature Feedback

not propagate to others. With the exception of Sun and DeJong [2005], there is also a lack of theoretical analysis of the efficacy of the various methods.

### 3 A PROBABILISTIC DISJUNCTION MODEL (PDM)

In this section, we define a stochastic model that generates the label  $y \in \{1, 2, \dots, k\}$  of any document  $d$ . The model makes use of an intermediate-level representation that, for concreteness, we think of as referring to topics.

Suppose we have a set of  $T$  “topics” as well as a procedure for representing any document as a convex combination  $\theta = (\theta_1, \dots, \theta_T)$  of these topics (so the  $\theta_t$  are nonnegative and sum to 1). The details of how this is done are irrelevant. We will assume that every topic  $t \in \{1, 2, \dots, T\}$  either has an associated label  $\ell(t) \in \{1, 2, \dots, k\}$  or has  $\ell(t) = ?$ . In the former case, the topic is a strong predictor of the corresponding label. In the latter case, the topic is ambiguous, for instance, an overly general topic. We will denote the set of predictive topics as  $P = \{t : \ell(t) \neq ?\}$  and we will assume that every document assigns non-zero probability to at least one predictive topic, that is,  $\sum_{t \in P} \theta_t > 0$ .

The *probabilistic disjunction model* is a generative process for the label of a document:

- Let  $\theta = (\theta_1, \dots, \theta_T)$  be the topic representation of the document.
- Pick a predictive topic at random: choose  $t \in P$  with probability proportional to  $\theta_t$ .
- The label of the document is  $\ell(t)$ .

Suppose there is no feature feedback; that is, the learner has access only to a collection of (document, label)

pairs. A reasonable objective, under the above stochastic model, is to find the assignment  $\ell : \{1, 2, \dots, T\} \rightarrow \{1, 2, \dots, k, ?\}$  that maximizes the likelihood of the data. But we can show that merely finding an assignment with non-zero likelihood is NP-hard.

**Theorem 1.** *The following problem is NP-complete: Given a collection of labeled documents, where each document is represented as a distribution over topics, and where  $k = 2$  (binary labels), find an assignment  $\ell : [T] \rightarrow \{0, 1, ?\}$  with non-zero likelihood.*

(Proof in Section A.1.) Feature feedback makes this intractability go away, as we will see next.

#### 3.1 Learning a PDM with Feature Feedback

The interactive labeling process works as follows:

- (a) The labeler gets a batch of (say) 10 documents.
- (b) For each document: he/she assigns it a label and chooses a predictive word (or maybe several words). This is then repeated until the budget for human interaction runs out.

The goal of the learner is to identify the correct mapping  $\ell : [T] \rightarrow \{1, 2, \dots, k, ?\}$ . A scheme for doing this is shown in Algorithm 1. Roughly, when the user tags a document with label  $y$  and identifies relevant words  $w_1, \dots, w_c$ , the algorithm picks a set of topics  $S \subseteq [T]$  triggered by these words and increments a counter  $n_{ty}$  for each  $t \in S$ . This  $n_{ty}$  counts how many times the user has suggested that topic  $t$  is predictive of label  $y$ .

The specific mechanism for choosing the set  $S$  based on the feedback, corresponding to the function `select-topics` in the pseudocode, is not relevant for the theoretical results we establish below. In our experimental work, we use the following strategy: given feedback words  $w_1, \dots, w_c$  for document  $x$ , obtain topic distributions for each of these words in the context of

document  $x$ ; call these  $p_1, \dots, p_c$  (distributions over  $T$  topics). Add topic  $t$  to the selected set  $S$  if the  $t$ th entry of  $(p_1 + \dots + p_c)/c$  exceeds a predefined threshold.

---

**Algorithm 1** Probabilistic Disjunction Model (PDM)

```

Input: Collection of unlabeled documents  $U$ 
Initialize:  $n_{ty} = 0, \forall t, y$ 
            Labeled data set  $L = \emptyset$ 
repeat
    Draw next batch  $B \subset U$  of documents at random
     $U = U \setminus B$ 
    for each document  $x \in B$  do
        Receive label  $y$ , relevant words  $w_1, \dots, w_c$ 
        Add  $(x, y)$  to  $L$ 
         $S = \text{select-topics}(x, w_1, \dots, w_c)$ 
        for  $t \in S$  do
             $n_{ty} = n_{ty} + 1$ 
        end for
    end for
    until budget runs out

```

---

**Assigning a label to each topic.** This is summarized in Algorithm 2. The total amount of feedback received for topic  $t$  is  $n_t = \sum_y n_{ty}$ . If this exceeds some fixed amount  $n_o$ , and moreover there is a specific label  $y$  for which  $n_{ty} \geq \lambda n_t$ , then we assign  $\hat{\ell}(t) = y$ . Here  $\lambda$  is a fixed fraction. In all other cases, we set  $\hat{\ell}(t) = ?$ .

**Labeling a new document.** This prediction rule is shown in Algorithm 3. Once topics are labeled, the estimated set of predictive topics is  $\hat{P} = \{t : \hat{\ell}(t) \neq ?\}$ . Let  $\theta$  be the topic distribution for the new document. The conditional probability that this document has label  $y$  is estimated as

$$\pi(y) = \frac{\sum_{t:\hat{\ell}(t)=y} \theta_t}{\sum_{t \in \hat{P}} \theta_t}.$$

### 3.2 Theoretical Guarantees

**Correctness of topic labeling.** In order to show that the topic labeling algorithm recovers the true labels  $\ell(t)$  with high probability, we do not need the full strength of the PDM assumption. What we require is that the topics selected by the user are not systematically misleading. On each round, the machine associates a set of user-selected topics  $S$  with a label  $y$ . Some of these associations may be spurious, for instance, due to polysemy that the user inadvertently overlooks. But the same spurious associations should not occur repeatedly.

To formalize this, first observe that the two sources of randomness in topic labeling are: (1) the random selection of documents for labeling, and (2) the possibly stochastic mechanism by which the human selects helpful words from a document.

---

**Algorithm 2** Topic labeling assignment (TLA)

```

Input:  $n_{ty} \forall t, y, \lambda, n_o$ 
for each topic  $t$  do
     $\hat{\ell}(t) = ?$ 
     $n_t = \sum_y n_{ty}$ 
    if  $n_t \geq n_o$  then
         $y = \arg \max_{y'} n_{ty'}$ 
        if  $n_{ty} \geq \lambda n_t$  then
             $\hat{\ell}(t) = y$ 
        end if
    end if
end for

```

---

**Algorithm 3** PDM prediction rule

```

Input: Topic representation  $\theta \in [0, 1]^T$  of document  $d$ 
Initialize:  $\pi = 0^k$ 
Label topics according to TLA (Algorithm 2)
for each topic  $t$  do
    if  $\hat{\ell}(t) \neq ?$  then
         $\pi(\hat{\ell}(t)) \leftarrow \pi(\hat{\ell}(t)) + \theta_t$ 
    end if
end for
Normalize  $\pi$  to sum to 1

```

---

**Assumption 1.** For any topic  $t$  and any label  $y \neq \ell(t)$ , if we pick a document at random, ask the human for the label and for helpful words, and look at the induced set of selected topics,

$$\Pr(\text{label} = y \mid \text{topic } t \text{ is selected}) \leq \lambda/2.$$

Meanwhile, for any predictive topic  $t \in P$ ,

$$\Pr(\text{label} = \ell(t) \mid \text{topic } t \text{ is selected}) \geq 2\lambda.$$

**Theorem 2.** Pick any  $0 < \delta < 1$ . Suppose Assumption 1 holds and that we set  $n_o \geq (6/\lambda) \ln(Tk)/\delta$ . Then with probability at least  $1 - \delta$ , for all  $t \in [T]$  with  $n_t \geq n_o$ , we have  $\hat{\ell}(t) = \ell(t)$ .

(Proof in Section A.2.)

**Label complexity.** In order to quantify the amount of feedback needed to recover the true labels  $\ell$ , we require that the user doesn't systematically avoid any informative topics, as follows.

**Assumption 2.** There is an absolute constant  $c_o$  for which the following holds. Pick any  $t, y$  such that  $\ell(t) = y$ . Then for any document with topic distribution  $\theta$  and label  $y$ , if we solicit feature feedback and look at the induced set of topics,

$$\Pr(\text{topic } t \text{ is selected}) \geq c_o \frac{\theta_t}{\sum_{t':\ell(t')=y} \theta_{t'}}.$$

Let  $\theta(x) = (\theta_1(x), \dots, \theta_T(x))$  be the topic distribution for any document  $x$ . We define the *prevalence* of a predictive topic  $t \in P$  as

$$\gamma_t = \mathbb{E}_x \left[ \frac{\theta_t(x)}{\sum_{t' \in P} \theta_{t'}(x)} \right],$$

where the expectation is over a uniform-random choice of  $x$  from the corpus. Roughly,  $\gamma_t$  tells us how common topic  $t$  is relative to other predictive topics, and thereby how easy it is to estimate  $\ell(t)$ .

**Theorem 3.** *Suppose documents are labeled according to the PDM process. Under Assumption 2, for any  $t \in P$ , the expected number of labels needed for  $\ell(t)$  to be set is at most  $n_o/(c_o \gamma_t)$ .*

(Proof in Section A.3.) For fixed constants  $\lambda$  and  $\delta$ , we need  $n_o = O(\ln Tk)$ . If all predictive topics are equally prevalent then they each have  $\gamma_t = 1/|P|$ . In this case, the number of rounds of interaction needed is  $O(|P| \ln(Tk))$ . This shows the benefit of feature feedback when only a small fraction of the topics are predictive (that is,  $|P| \ll T$ ).

## 4 LEARNING LINEAR SEPARATORS WITH FEATURE FEEDBACK

We now study feature feedback in the setting where the goal is to learn a linear classifier by minimizing a loss function and a regularization penalty. Given a data set  $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathcal{Y}$ , the optimization is:

$$\hat{w} = \arg \min_w \frac{1}{n} \sum_{i=1}^n \ell(w \cdot x_i, y_i) + \lambda \|w\|^2,$$

where  $\ell(\cdot)$  is a loss function and  $\|\cdot\|$  is some norm. For SVMs, for instance,  $\ell$  is the hinge loss and  $\|\cdot\|$  is the 2-norm.

We propose a simple scheme for incorporating information about relevant features: reduce the regularization along those specific dimensions. To achieve this, we take the regularization norm  $\|\cdot\|$  to be a *Mahalanobis* norm, given by a  $p \times p$  positive definite matrix  $A$ :

$$\|x\|_A = \sqrt{x^T A x} = \|A^{1/2} x\|_2.$$

In the absence of feature feedback,  $A$  is the identity matrix  $I_p$ , giving the 2-norm. But if we find that features  $R \subset [p]$  are relevant, we downweight the diagonal matrix in those dimensions: we set  $A_{jj} = 1/c$  for relevant features  $j$  and  $A_{jj} = 1$  otherwise, for some  $c > 1$ . In spirit, this regularization reweighting is analogous to increasing the prior on these features in a Bayesian model, as was done in Settles [2011].

We next study the statistical benefit of this estimator.

### 4.1 Improved Generalization Error Bounds

Let's start with a generalization bound for learning linear classifiers chosen from some set  $\mathcal{F}$ . Write the empirical loss function as

$$\widehat{\mathcal{L}}(w) = \frac{1}{n} \sum_{i=1}^n \ell(w \cdot x_i, y_i)$$

(regularization is incorporated by restricting  $\mathcal{F}$  to vectors of bounded norm). When the training data  $(x_i, y_i)$  comes i.i.d. from an (unknown) underlying distribution, the following seminal result shows the relation of  $\widehat{\mathcal{L}}(w)$  to the true loss  $\mathcal{L}(w) = \mathbb{E}_{x,y} \ell(w \cdot x, y)$ :

**Theorem 4.** [Bartlett and Mendelson, 2003] *Suppose the loss function  $\ell$  is Lipschitz in its first argument and is upper-bounded by a constant  $M_\ell$ . Then for any  $\delta > 0$ , with probability  $\geq 1 - \delta$  over the choice of data,*

$$\forall f \in \mathcal{F}: \quad \mathcal{L}(f) \leq \widehat{\mathcal{L}}(f) + 2R_n(\mathcal{F}) + M_\ell \sqrt{\frac{\log 1/\delta}{2n}},$$

where  $R_n(\mathcal{F})$  is the Rademacher complexity of  $\mathcal{F}$ .

The key term here is  $R_n(\mathcal{F})$ . In our setup, let  $w^*$  be a sparse target classifier of interest and define a feature as being relevant if it is set in  $w^*$ . Using a powerful result of [Kakade et al., 2009], we obtain the following.

**Theorem 5.** *Let  $R = \{j \in [p] : w_i^* \neq 0\}$  denote the relevant features of  $w^*$ .*

- *We can write any  $x$  in terms of its relevant and other components,  $x = (x_R, x_o)$ .*
- *Let  $A$  be the diagonal matrix whose  $j$ th entry is  $1/c$  if  $j \in R$  and 1 otherwise.*

*Then, for the family of linear separators  $\mathcal{F} = \{w : \|w\|_A \leq \|w^*\|_A\}$ , we have*

$$R_n(\mathcal{F}) \leq \|w^*\|_2 \cdot \max_{x \in \mathcal{X}} \sqrt{\left( \frac{1}{c} \|x_o\|_2^2 + \|x_R\|_2^2 \right)} \sqrt{\frac{2}{n}}.$$

(Proof in Section B.1.) In situations where the  $x_o$  (the irrelevant portion of the data) has significant norm, this downweighting by a factor of  $c$  substantially reduces the generalization error bound. Note that this result requires all relevant features to be revealed in advance. It remains an open problem to characterize the benefit when relevant features are gradually disclosed during rounds of interaction.

### 4.2 Practical Linear Models with Feature Feedback

Given training data  $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathcal{Y}$  consider the SVM problem with our Mahalanobis regularizer:

---

**Algorithm 4** SVM with feature feedback (SVM-FF)

---

```

Input:  $c < 1$ , unlabeled data set  $U$ 
Initialize:  $L = \emptyset, A = I_p$ 
repeat
    Draw next batch  $B \subset U$  of documents
     $U = U \setminus B$ 
    for each document  $x \in B$  do
        Receive label  $y$ , words  $s$ 
        Add  $(x, y)$  to  $L$ 
        for  $j \in s$  do
             $A_{jj} = c$ 
        end for
        Train linear SVM on  $\{(A^{-1/2}x, y) : (x, y) \in L\}$ 
    end for
until budget runs out

```

---

**Algorithm 5** Bootstrap PDM

---

```

Input: Unlabeled data set  $U, \tau_0$  (optionally,  $c < 1$ )
Initialize:  $L = \emptyset$  (optionally,  $A = I_p$ )
repeat
    Draw next batch  $B \subset U$  of documents
     $L = L \cup B; U = U \setminus B$ 
    Train PDM (Algorithm 1) on  $L$ 
    (optionally, update  $A$  as in Algorithm 4)
    for each document  $x \in U$  do
         $I = \emptyset$  (documents with inferred labels)
        Predict  $\pi(\cdot)$  over labels according to Algorithm 3
        Predict  $\hat{y} = \arg \max_{y' \in \{1, \dots, k\}} \pi(y')$ 
        if  $\pi(\hat{y}) \geq \tau_0$  then
            Add  $(x, \hat{y})$  to  $I$ 
        end if
    end for
    Train any classifier on  $\{(x, y) : (x, y) \in L \cup I\}$ 
    (optionally, train linear SVM as in Algorithm 4)
until budget runs out

```

---

$$\begin{aligned}
 & \underset{w}{\text{minimize}} \quad \frac{1}{2} \|w\|_A^2 + C \sum_{i=1}^N \xi_i \\
 & \text{subject to} \quad \xi_i \geq 0, \quad y_i(x_i^T w + b) \geq 1 - \xi_i, \quad \forall i.
 \end{aligned}$$

A straightforward derivation shows the following.

**Lemma 6.** *Pick any positive definite  $p \times p$  matrix  $A$ . Then, learning a linear SVM on instances  $\{(x_i, y_i)\}_{i=1}^n$  with Mahalanobis regularizer  $\|w\|_A$  is equivalent to learning a linear SVM on modified instances  $\{(A^{-1/2}x_i, y_i)\}_{i=1}^n$  with  $\|w\|_2$  regularization.*

(Proof in Section C.) An SVM algorithm with feature feedback (SVM-FF) is given in Algorithm 4. For each supplied feature, the corresponding diagonal entries of  $A$  are set to a particular value  $c < 1$  and every labeled and unlabeled example is weighted by  $A^{-1/2}$ . Then, a standard linear SVM is trained on the weighted labeled instances.

## 5 BOOTSTRAPPING THE PDM

The feedback in the regularization approach is explicit: the regularization will only be applied to features that the labeler selects. Let's return to the "filibuster"- "politics" example in the introduction. Even though the word "filibuster" is a good predictor for "politics" it is a fairly uncommon word. Hence, not that many documents will be affected by reducing the regularization on it. On the other hand, vague feature feedback facilitated by the PDM is richer: feedback on "filibuster" propagates to other words in the same topic. To incorporate vague feedback into a linear classifier, we introduce the *bootstrapped* PDM (Algorithm 5). Given a labeled data set  $L$  and an unlabeled data set  $U$ , the algorithm fits a PDM to  $L$  and uses this PDM to predict on  $U$ . It then infers the labels of a set  $I \subseteq U$  of data points for which it is confident. We say that the PDM is confident on an instance  $x$  if its prediction  $\hat{y}$  has estimated conditional probability  $\pi(\hat{y}) \geq \tau_0$  (recall the notation of Algorithm 3), where  $\tau_0$  is a parameter to be set. One can then train any classifier on  $L \cup I$ . If the classifier of choice is a linear SVM, one can apply the mixed regularization, by multiplying every example by  $A^{-1/2}$  and training a linear SVM on this weighted data set of labeled and inferred points.

## 6 EXPERIMENTS

We conducted experiments on the following 6 benchmark text categorization data sets. **20 NewsGroups**: Set of approximately 20,000 documents, partitioned evenly across 20 newsgroups, containing postings about politics, sports, technology, religion, science etc. **Reuters-21578**: Another widely used collection for text categorization research. Documents with less than or with more than one label were eliminated, resulting in **R8** (8 classes) and **R52** (52 classes). **webkb**: Data set that contains web pages collected from computer science departments of various universities. **cade**: Web pages from the CADE Web Directory, which points to Brazilian web pages classified by human experts in 12 classes, including services, education, sciences, sports, culture etc. **ohsumed**: Medical abstracts from the MeSH (Medical Subject Headings) data set, belonging to 23 cardiovascular disease categories. For further details on the data sets, see section D.1 of the Appendix. The first five data sets were already processed [Cardoso-Cachopo, 2007]; we processed **ohsumed** in the same manner (stemming, removal of stop words and words shorter than two characters). As we are interested in single label documents, we only kept data points that had only one label. For each document we obtained its tf-idf and topic representations. For the latter we trained a Latent Dirichlet Allocation model using the

collapsed Gibbs sampler [Griffiths and Steyvers, 2004]. The number of topics was 10 times the number of classes in each data set.

**Oracle features.** To simulate the labeler’s feedback, we first generated a list of *oracle* features for each class as follows. We first trained a logistic regression classifier with  $\ell_1$  regularization and took all the feature weights that were positive. We then looked at the level of correlation between these features and the class labels. Specifically, for various thresholds  $\alpha$ , we considered feature  $j$  as feedback for class  $k$  if  $P(k|j)$ , the conditional probability of label  $k$  given the presence in the document of word  $j$ , was at least  $\alpha$ . We then tested our models for various values of  $\alpha$ . Feature feedback on a document applied if it contained any of the words in the list of its label. An example of feature feedback for the **20ng** dataset using the PDM is shown in figure 4 in the appendix.

**Experimental setup.** We compared our models to a linear SVM without feedback. To choose the cost  $C$  of all SVM classifiers, we only tuned the SVM without feedback by optimizing the macro- $F_1$  score on the grid  $\{1, 10, 100, 1000\}$ . We then set  $C$  for the SVM-FF and bootstrap PDM models to that value. On the first few batch iterations we used 2-fold cross validation and continued with 5-fold in later iterations. We set the rest of the parameters for PDM, SVM-FF, and bootstrap PDM as follows:  $\lambda = \frac{1}{10}$ ,  $n_o = 2$ ,  $c = \frac{1}{20}$  and  $\tau_0 = .75$ .

**Discussion of simulation results.** Figures 2 (a-c) show learning curves for the first 500 data points for each training data set, divided into 20 batches. For each batch iteration, we report macro- $F_1$  score on the test set. (See D.2 for a more detailed exposition of the experimental results.) Across the board, we find that feedback on a few predictive words helps significantly. To get a feel of the *amount* of feature feedback see figures 11- 12 in D.2. Vague feature feedback (PDM, bootstrap PDM) is particularly helpful when the labeled data set is small. Generous feature feedback (i.e.  $\alpha \geq .5$ ) helps fast convergence when data are scarce but has a somewhat adverse effect when plenty of labeled samples are available. However, this improves for  $\alpha \geq .9$ . Interestingly, in addition to its superior performance, SVM-FF produces a solution that is much sparser than that of the SVM, as seen in figure 2d. This makes sense intuitively, as feature feedback helps the learning algorithm to focus on important dimensions.

**Small vs large data regimes.** The simulation results illustrate that the benefits of feature feedback diminish asymptotically. We note that since we are learning a linear classifier, in the limit of enough labeled data, we can simply run SVM. Also, the degree

of regularization in the SVM-FF can be adjusted so that  $c \rightarrow 1$  as the sample grows. Hence, our methods are well suited to the fairly common situation where the amount of labeled data is limited.

**Human experiment.** To get a sense of the feature feedback that humans tend to provide and to quantify the difference in the benefits of a *selected* feature *vs* a *random* feature, we conducted a small human study involving 5 annotators. We considered a subset of the **20ng** data set that included points with classes *talk.politics.mideast*, *comp.graphics*, *sci.med*, *rec.autos* and *misc.forsale*. The annotators provided the labels of a randomly chosen set of 50 points along with a number of features via an interface. (See D.3 for details). For class  $k$ , call  $S_k$ ,  $N_k$  the set of features that annotators selected and did not select, respectively. In table 1 we show  $\bar{p}_{S_k} = \frac{1}{|S_k|} \sum_{j \in S_k} P(k|j)$  and  $\bar{p}_{N_k} = \frac{1}{|N_k|} \sum_{j \in N_k} P(k|j)$ , where the  $P(k|j)$ ’s are the conditional probabilities described earlier.

Table 1: Results of Human Experiment

|                              | $\bar{p}_{S_k}$ | $\bar{p}_{N_k}$ |
|------------------------------|-----------------|-----------------|
| <i>misc.forsale</i>          | 0.63            | 0.76            |
| <i>rec.autos</i>             | 0.95            | 0.82            |
| <i>sci.med</i>               | 0.96            | 0.78            |
| <i>comp.graphics</i>         | 0.83            | 0.66            |
| <i>talk.politics.mideast</i> | 0.98            | 0.74            |

Note that  $\bar{p}_{S_k}$  is smaller than  $\bar{p}_{N_k}$  only for the class *misc.forsale* because some annotators confused documents about items for sale with documents with class *comp.graphics* and *rec.autos*. This is not a surprising effect and we expect to diminish with more labeled data and with a larger pool of annotators. Across the board, we find that humans tend to provide words that are highly predictive of the label.

## Acknowledgements

The authors are grateful to National Science Foundation for support under grant IIS-1162581, for feedback from the anonymous reviewers, and for help with the human experiment from Julaiti Alafate, Zack Lipton, Christopher Tosh, and Sharad Vikram.

## References

- P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.
- S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition

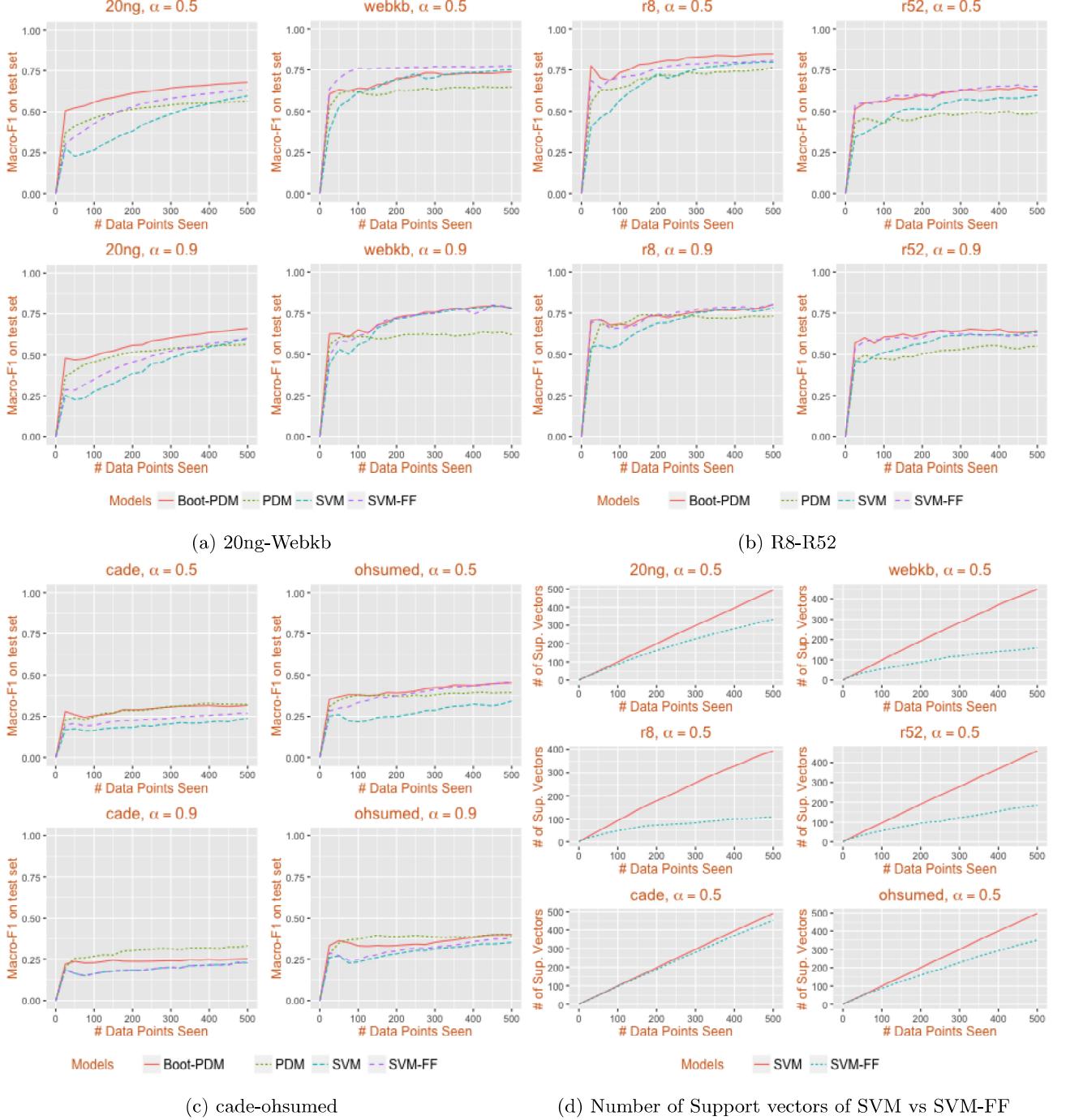


Figure 2: (a) to (c): Learning Curves at Different Values of  $\alpha$ . (d): Number of Support Vectors.

- with humans in the loop. In *European Conference on Computer Vision*, 2010.
- A. Cardoso-Cachopo. *Improving methods for single-label text categorization*. PhD thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa, 2007.
- W. Croft and R. Das. Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, pages 349–368, 1990.
- S. Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781, 2011.
- A. Dayanik, D. Lewis, D. Madigan, V. Menkov, and A. Genkin. Constructing informative prior distributions from domain knowledge in text classification. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 493–500, 2006.
- J. Donahue and K. Grauman. Annotator rationales for visual recognition. In *2011 IEEE International Conference on Computer Vision*, pages 1395–1402, 2011.
- G. Druck, G. Mann, and A. McCallum. Reducing annotation effort using generalized expectation criteria (technical report 2007-62). *University of Massachusetts, Amherst*, 2007.
- G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of ACM Special Interest Group on Information Retrieval*, 2008.
- T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.
- S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, pages 793–800, 2009.
- D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.
- P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining*, pages 483–486, 2004.
- P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. An expected utility approach to active feature-value acquisition. In *Fifth IEEE International Conference on Data Mining*, pages 4–pp, 2005.
- H. Raghavan and J. Allan. An interactive algorithm for asking and incorporating feature feedback into support vector machines. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 79–86. ACM, 2007.
- H. Raghavan, O. Madani, and R. Jones. Interactive feature selection. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 841–846, 2005.
- H. Raghavan, O. Madani, and R. Jones. Active learning with feedback on features and instances. *The Journal of Machine Learning Research*, 7:1655–1686, 2006.
- R. Schapire, M. Rochery, M. Rahim, and N. Gupta. Incorporating prior knowledge into boosting. In *Proceedings of the 19th International Conference on Machine Learning*, volume 2, pages 538–545, 2002.
- B. Settles. Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances. In *Empirical Methods in Natural Language Processing*, 2011.
- B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.
- V. Sindhwani, P. Melville, and R. Lawrence. Uncertainty sampling and transductive experimental design for active dual supervision. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 953–960, 2009.
- K. Small, B. Wallace, T. Trikalinos, and C. Brodley. The constrained weight space svm: learning with ranked features. In *Proceedings of the 28th International Conference on Machine Learning*, pages 865–872, 2011.
- Q. Sun and G. DeJong. Explanation-augmented svm: an approach to incorporating domain knowledge into svm learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 864–871, 2005.
- X. Wu and R. Srihari. Incorporating prior knowledge with weighted margin support vector machines. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 326–333, 2004.
- O. Zaidan and J. Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 31–40, 2008.
- O. Zaidan, J. Eisner, and C. Piatko. Using “annotator rationales” to improve machine learning for text categorization. In *Human Language Technology*

*Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings,*  
pages 260–267, 2007.