Growing a List

Benjamin Letham
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139
bletham@mit.edu

Cynthia Rudin
MIT Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139
rudin@mit.edu

Katherine A. Heller
Center for Cognitive Neuroscience
Statistical Science
Duke University
Durham, NC 27708
kheller@gmail.com

Abstract

It is easy to find expert knowledge on the Internet on almost any topic, but obtaining a complete overview of a given topic is not always easy: Information can be scattered across many sources and must be aggregated to be useful. We introduce a method for intelligently growing a list of relevant items, starting from a small seed of examples. Our algorithm takes advantage of the wisdom of the crowd, in the sense that there are many experts who post lists of things on the Internet. We use a collection of simple machine learning components to find these experts and aggregate their lists to produce a single complete and meaningful list. We use experiments with gold standards and open-ended experiments without gold standards to show that our method significantly outperforms the state of the art. Our method uses the clustering algorithm Bayesian Sets even when its underlying independence assumption is violated, and we provide a theoretical generalization bound to motivate its use.

1 Introduction

We aim to use the collective intelligence of the world's experts to grow a list of useful information on any given topic. To do this, we aggregate knowledge from many experts' online guides in order to create a central, authoritative source list. We focus on the task of open-ended list aggregation, inspired by the collective intelligence problem of finding all planned events in a city. There are many online "experts" that list Boston events, such as Boston.com or Yelp, however these lists are incomplete. As an example of the difficulties caused by information fragmentation, traffic in parts of greater Boston can be particularly bad when there is a large public event such as a street festival or fundraising walk. There are a number of lists of Boston events, but they are all incomplete. Even though these events are planned well in advance, the lack of a central list of events makes it hard to avoid traffic jams, and the number of online sources makes it difficult to compile a complete list manually.

As the amount of information on the Internet continues to grow, it becomes increasingly important to be able to compile information automatically in a fairly complete way, for any given domain. The development of general methods that automatically aggregate this kind of collective knowledge is a vital area of current research, with the potential to positively impact the spread of useful information to users across the Internet.

Our contribution in this paper is a real system for growing lists of relevant items from a small "seed" of examples by aggregating information across many internet experts. We provide an objective evaluation of our method to show that it performs well on a wide variety of list growing tasks, and significantly outperforms existing methods. We provide some theoretical motivation by giving bounds for the Bayesian Sets algorithm used within our algorithm. None of the components of our method are particularly complicated; the value of our work lies in combining these simple ingredients in the right way to solve a real problem.

There are two existing methods for growing a list of items related to a user-specified seed. The problem was introduced ten years ago on a large scale by Google Sets, which is accessible via Google Spreadsheet. We also compare to a more recent online system called Boo!Wa! (http://boowa.com), which is similar in concept to Google Sets. In our experiments, we found that Boo!Wa! is a substantial advance above Google Sets, and the algorithm introduced here is a similarly sized leap in technology above Boo!Wa!. In a set of 50 experiments shown in Section 4, the lower 25th percentile of our performance was better than the median performance of both Google Sets and Boo!Wa!, both in Precision@5 and Precision@20. More generally, our work builds on "search" and other work in information retrieval. Search engines locate documents containing relevant information, but to produce a list one would generally need to look through the webpages and aggregate the information manually. We build on the speed of search, but do the aggregation automatically and in a much more complete way than a single search.

In the supplementary material, we provide: additional details on the algorithm implementation, the results for each gold standard experiment, three additional open-ended experiments, an additional generalization bound, and proofs of the theoretical results.

Algorithm 1 Outline of the list growing algorithm

```
Input: A list of seed items
Output: A ranked list of new items related to the seed items
for as many iterations as desired do
  for each pair of seed items do
    Source discovery: Find all sites containing both items
    for each source site do
      List extraction: Find all items on the site represented similarly to the seed items
    end for
  end for
  for each discovered item do
    Feature space: Construct a binary feature vector of domains where the item is found
    Ranking: Score the item according to the seed using Bayesian Sets
  end for
  Implicit feedback: Add the highest-ranked non-seed item to the seed
end for
```

2 Algorithm

Algorithm 1 gives an outline of the list growing algorithm, which we now discuss in detail.

Source discovery: We begin by using the seed items to locate sites on the Internet that serve as expert sources for other relevant items. We use a combinatorial search strategy that relies on the assumption that a site containing at least two of the seed items likely contains other items of interest. Specifically, for every pair of seed items, we search for all websites that contain both of the items; this step takes advantage of the speed of "search."

List extraction: The output of the combinatorial search is a list of source sites, each of which contains at least two seed items. We then extract all of the new items from each of these sites. Here our strategy relies on the assumption that human experts organize information on the Internet using HTML tags. For each site found with the combinatorial search, we look for HTML tags around the seed items. We then find the largest set of HTML tags that are common to both seed items, for this site, and extract all items on the page that use the same HTML tags. Because we allow any HTML tags, including generic ones like <b> and <a>, the lists we recover can be noisy. When we combine the lists together, we use a clustering algorithm to ensure that the noise is pushed to the bottom of the list.

Feature Space: At this point the algorithm has discovered a collection of lists, each from a different source. We now combine these lists so that the most relevant information is on the top of the final, merged list. To determine which of the discovered items are relevant, we construct a feature space in which to compare them to the seed items. Specifically, for each discovered item x, we construct a binary feature vector where each feature j corresponds to an internet domain (like boston.com or mit.edu), and $x_j = 1$ if item x can be found on internet domain j. This set of internet domains is found using a search engine with the item as the query. Related items should be found on a set of mainly overlapping domains, so we determine relevance by looking for items that cluster well with the seed items in the feature space.

Ranking: The Bayesian Sets algorithm (Ghahramani and Heller, 2005) is a clustering algorithm based on a probabilistic model for the feature space. Specifically, we suppose that each feature (in general, x_j) is a Bernoulli random variable with probability θ_j of success: $x_j \sim \text{Bern}(\theta_j)$. Following the typical Bayesian practice, we assign a Beta prior to the probability of success: $\theta_j \sim \text{Beta}(\alpha_j, \beta_j)$. Bayesian Sets assigns a score $f(x)$ to each item x by comparing the likelihood that x and the seed $S = \{x^1, \ldots, x^m\}$ were generated by the same distribution to the likelihood they are independent:

$$f(x) := \log \frac{p(x, S)}{p(x)p(S)}. \tag{1}$$

Suppose there are N features: $x \in \{0,1\}^N$. Because of the Bernoulli-Beta conjugacy, Ghahramani and Heller (2005) show that (1) has an analytical form under the assumption of independent features. However, the score given in Ghahramani and Heller (2005) can be arbitrarily large as m (the number of seed examples) increases. We prefer a normalized score for the purpose of the generalization bound, and so we use the following scoring function which differs from that in Ghahramani and Heller (2005) only by constant factors and normalization:

$$\begin{aligned} f_S(x) :=& \frac{1}{Z(m)} \sum_{j=1}^{N} x_j \log \frac{\alpha_j + \sum_{s=1}^m x_j^s}{\alpha_j} \\ &+ (1 - x_j) \log \frac{\beta_j + m - \sum_{s=1}^m x_j^s}{\beta_j}, \end{aligned} \tag{2}$$

where

$$Z(m) := N \log \left(\frac{\gamma_{\min} + m}{\gamma_{\min}} \right)$$

and $\gamma_{\min} := \min_j \min\{\alpha_j, \beta_j\}$ is the weakest prior hyperparameter. It is easy to show that $f_S(x) \in [0,1]$. Given the seed and the prior, (2) is linear in x, and can be formulated as a single matrix multiplication. When items are scored using Bayesian Sets, the items that were most likely to have been generated by the same distribution as the seed items are put high on the list.

Feedback: Once the lists have been combined, we continue the discovery process by expanding the seed. A natural, unsupervised way of expanding the seed is to add the highest ranked non-seed item into the seed. Though not done here, one could also use a domain expert or even crowdsourcing to quickly scan the top ranked items and manually expand the seed from the discovered items. Then the process starts again; we do a combinatorial search for websites containing all pairs with the new seed item(s), extract possible new items from the websites, etc. We continue this process for as many iterations as we desire.

Further implementation details are available in the supplementary material.

3 Theoretical Results

The derivation for Bayesian Sets assumes independent features. In this application, features are internet domains, which are almost certainly correlated. Because Bayesian sets is the core of our method, we motivate its use in this application by showing that even in the presence of arbitrary dependence among features, prediction ability can be guaranteed as the sample size increases. We consider an arbitrary distribution from which the seed S is drawn, and prove that as long as there are a sufficient number of items, x will on expectation score highly as long as it is from the same distribution as the seed S. Specifically, we provide a lower bound for $\mathbb{E}_x\left[f_S(x)\right]$ that shows that the expected score of x is close to the score of S with high probability.

Theorem 1. *Suppose $x^1, \dots, x^m$ are sampled independently from the same distribution $\mathcal{D}$. Let $p_{\min} = \min_j \min\{p_j, 1-p_j\}$ be the probability of the rarest feature. For all $p_{\min} > 0$, $\gamma_{\min} > 0$ and $m \geq 2$, with probability at least $1-\delta$ on the draw of the training set $S = \{x^1, \dots, x^m\}$,*

$$\mathbb{E}_{x\sim\mathcal{D}}\left[f_S(x)\right] \geq \frac{1}{m}\sum_{s=1}^{m} f_S(x^s) - \sqrt{\frac{1}{2m\delta} + \frac{6}{g(m)\delta} + O\left(\frac{1}{m^2 \log m}\right)},$$

where,

$$g(m) := \log\left(\frac{\gamma_{\min} + m - 1}{\gamma_{\min}}\right)\left(\gamma_{\min} + (m-1)p_{\min}\right)$$

The proof technique involves showing that Bayesian Sets is a "stable" algorithm, in the sense of "pointwise hypothesis stability" (Bousquet and Elisseeff, 2002). We show that the Bayesian Sets score is not too sensitive to perturbations in the seed set. Specifically, when an item is removed from the seed, the average change in score is bounded by a quantity that decays as $\frac{1}{m \log m}$. This stability allows us to apply a generalization bound from Bousquet and Elisseeff (2002). The proof of pointwise hypothesis stability is in the supplementary material.

The two quantities with the most direct influence on the bound are $\gamma_{\min}$ and $p_{\min}$. We show in the supplementary material that for $p_{\min}$ small relative to $\gamma_{\min}$, the bound improves as $\gamma_{\min}$ increases (a stronger prior). This suggests that a strong prior improves stability when learning data with rare features. As $p_{\min}$ decreases, the bound becomes looser, suggesting that datasets with rare features will be harder to learn and will be more prone to errors.

It is useful to note that the bound does not depend on the number of features N, as it would if we considered Bayesian Sets to simply be a linear classifier in N dimensions or if we used a straightforward application of Hoeffding's inequality and the union bound. Although a Hoeffding's inequality-based bound does provide a tighter dependence on δ due to the use here of Chebyshev's inequality rather than Hoeffding's inequality (for example, a Hoeffding-based bound is given in the supplementary material), the bound depends on N which in this application is the number of internet domains, and is thus extremely large. The fact that the bound in Theorem 1 is independent of N provides motivation for using Bayesian Sets on very large scale problems, even when the feature independence assumption does not hold.

The gap between the expected score of x and the (empirical) score of the seed goes to zero as $\frac{1}{\sqrt{m}}$. Thus when the seed is sufficiently large, regardless of the distribution over relevant items, we can be assured that the relevant items generally have high scores.

4 Experiments

We demonstrate and evaluate the algorithm with two sets of experiments. In the first set of experiments, we provide an objective comparison between our method, Google Sets, and Boo!Wa! using a randomly selected collection of list growing problems for which there exist gold standard lists. The true value of our work lies in the ability to construct lists for which there are not gold standards, so in a second set of experiments we demonstrate the algorithm's performance on more realistic, open-ended list growing problems. For all experiments, the steps and parameter settings of the algorithm were exactly the same and completely unsupervised other than specifying two seed items.

4.1 Wikipedia Gold Standard Lists

An objective evaluation of our method requires a set of problems for which gold standard lists are available. The "List of ..." articles on Wikipedia form a large corpus of potential gold standard lists that cover a wide variety of topics. We limited our experiments to the "featured lists," which are a collection of over 2,000 Wikipedia lists that meet certain minimum quality criteria. We required the lists used in our experiments to have at least 20 items, and excluded any lists of numbers (such as dates or sports scores). We created a random sample of list growing problems by randomly selecting 50 Wikipedia lists that met the above requirements. The selected lists covered a wide range of topics, including, for example, "storms in the 2005 Atlantic hurricane season," "current sovereign monarchs," "tallest buildings in New Orleans," "X-Men video games," and "Pittsburgh Steelers first-round draft picks." We treated the Wikipedia list as the gold standard for the associated list growing problem. We give the names of all of the selected lists in the supplementary material.

For each of the 50 list growing problems, we randomly selected two list items from the gold standard to form a seed. We used the seed as an input to our algorithm, and ran one iteration. We used the same seed as an input to Google Sets and Boo!Wa!. We compare the lists returned by our method, Google Sets, and Boo!Wa! to the gold standard list by computing the precision at two points in the rankings: Precision@5 and Precision@20. This measures the fraction of items up to and including that point in the ranking that are found on the gold standard list.

In Figures 1 and 2 we show boxplots of the precision results across all 50 gold standard experiments. For both Google Sets and Boo!Wa!, the median precision at both 5 and 20 was 0. Our method performed significantly better, with median precision of 0.4 and 0.425 at 5 and 20 respectively. For our algorithm, the lower quartile for the precision was 0.2 and 0.15 for 5 and 20 respectively, whereas this was 0 for Google Sets and Boo!Wa! at both precision levels. Our method returned at least one relevant result in the top 5 for 82% of the experiments, whereas Google Sets and Boo!Wa! returned at least one relevant result in the top 5 for only 22% and 38% of experiments, respectively.

The supplementary material gives a list of the Precision@5 and Precision@20 values for each of the Wikipedia gold standard experiments.

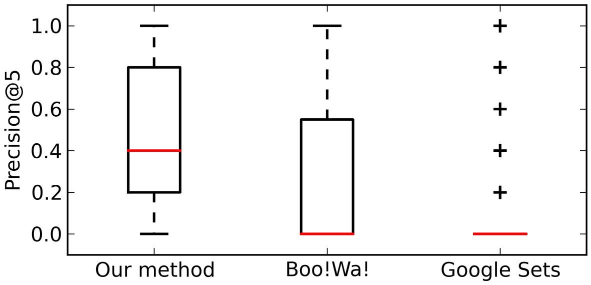

Figure 1: Precision@5 across all 50 list growing problems sampled from Wikipedia. The median is indicated in red.

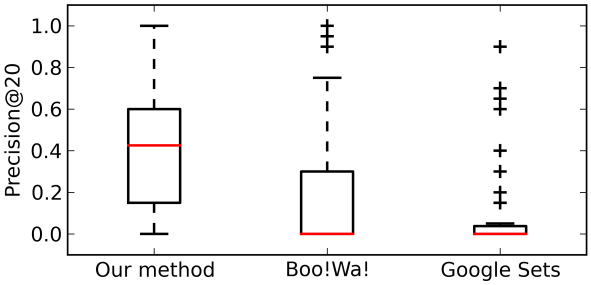

Figure 2: Precision@20 across all 50 list growing problems sampled from Wikipedia.

There are some flaws with using Wikipedia lists as gold standards in these experiments. First, the gold standards are available online and could potentially be pulled directly without requiring any aggregation of experts across different sites. However, all three methods had access to the gold standards and the experiments did not favor any particular method, thus the comparison is meaningful. A more interesting experiment is one that necessitates aggregation of experts across different sites; these experiments are given in Section 4.2. Second, these results are only accurate insofar as the Wikipedia gold standard lists are complete. We limited our experiments to "featured lists" to have the best possible gold standards. A truly objective comparison of methods requires both randomly selected list problems and gold standards, and the Wikipedia lists, while imperfect, provide a useful evaluation.

4.2 Open-Ended Experiments

It is somewhat artificial to replicate gold standard lists that are already on the Internet. In this set of experiments we demonstrate our method's performance on more realistic, open-ended list growing problems. For these problems gold standard lists are not available, and it is essential for the algorithm to aggregate results across many experts. We focus on two list growing problems: Boston events and Jewish foods. In the supplementary material we provide 3 additional open-ended list growing problems: smartphone apps, politicians, and machine learning conferences.

4.2.1 Boston Events

In this experiment, the seed items were two Boston events: "Boston arts festival" and "Boston harborfest." We ran the algorithm for 5 iterations, yielding 3,090 items. Figure 3 shows the top 50 ranked items, together with the source site where they were discovered. There is no gold standard list to compare to directly, but the results are overwhelmingly actual Boston events. The events were aggregated across a variety of expert sources, including event sites, blogs, travel guides, and hotel pages. Figure 4 shows the full set of results returned from Google Sets with the same two events as the seed. Not only is the list very short, but it does not contain any actual Boston events. Boo!Wa! was unable to return any results for this seed.

4.2.2 Jewish Foods

In this experiment, the seed items were two Jewish foods: "Challah" and "Knishes." Although there are lists of foods that are typically found in Jewish cuisine, there is variety across lists and no authoritative definition of what is or is not a Jewish food. We completed 5 iterations of the algorithm, yielding 8,748 items. Figure 5 shows the top 50 ranked items, together with their source sites. Almost all of the items are closely related to Jewish cuisine. The items on our list came from a wide variety of expert sources that include blogs, informational sites, bakery sites, recipe sites, dictionaries, and restaurant menus. In fact, the top 100 most highly ranked items came from a total of 52 unique sites. In Figure 6, we show the complete set of results returned from Google Sets for the same seed of Jewish foods. Although the results are foods, they are not closely related to Jewish cuisine. Boo!Wa! was unable to return any results for this seed.

Boston events
Boston arts festival
Boston harborfest
Whats going this month
Interview with ann scott
Studio view with dannyo
Tony savarino
Artwalk 2011
Greater boston convention visitors bureau
Cambridge chamber of commerce
Boston tours
3 county fairground
Boston massacre

Figure 4: Google Sets results for the Boston events experiment (seed italicized).

5 Related Work

There is a substantial body of work in areas or tasks related to the one which we have presented, which we can only briefly review here. There are a number of papers on various aspects of "set expansion," often for completing lists of entities from structured lists, like those extracted from Wikipedia (Sarmento et al., 2007), using rules from natural language processing or topic models (Tran et al., 2010; Sadamitsu et al., 2011), or from opinion corpora (Zhang and Liu, 2011). The task we explore here is *web-based set expansion* (see, for example, Jindal and Roth, 2011) and methods developed for other set expansion tasks are not directly applicable.

Item	**Source**
^{0}Boston arts festival	(original seed)
^{3}Cambridge river festival	bizbash.com/bostons_top_100_events/boston/story/21513/
^{0}Boston harborfest *harborfest*	(original seed)
^{1}Boston chowderfest	celebrateboston.com/events.htm
^{4}Berklee beantown jazz festival *the berklee beantown jazz festival,* *berklee bean town jazz festival*	pbase.com/caseus/arts&view=tree
^{2}Chinatown main street festival *www.chinatownmainstreet.org*	blog.charlesgaterealty.com/365-things/?Tag=Boston%20life
4th of july boston pops concert & fireworks display *boston 4th of july fireworks & concert*	travel2boston.us/boston-harborfest-30th-anniversary-[...]
Boston common frog pond *ice skating on boston common frog pond*	bostonmamas.com/2009/06/
First night boston	what-is-there-to-do.com/Boston/Festivals.aspx
Boston dragon boat festival *hong kong dragon boat festival of boston* *dragon boat festival of boston*	pbase.com/caseus/arts&view=tree
Boston tea party re enactment	ef.com/ia/destinations/united-states/boston/student-life/
Christopher columbus waterfront park	bostonmamas.com/2009/09/
Jimmy fund scooper bowl	bizbash.com/bostons_top_100_events/boston/story/21513/
Opening our doors day	ef.com/ia/destinations/united-states/boston/student-life/
Oktoberfest harvard square & harpoon brewery	sheratonbostonhotel.com/boston-festivals-and-events
August moon festival	ef.com/ia/destinations/united-states/boston/student-life/
Annual boston wine festival	worldtravelguide.net/boston/events
Cambridge carnival	soulofamerica.com/boston-events.phtml
Regattabar	berklee.edu/events/summer/
Arts on the arcade	berklee.edu/events/summer/
Franklin park zoo	hotels-rates.com/hotels_reservations/property/27353/
Faneuil hall annual tree lighting ceremony	ef.com/ia/destinations/united-states/boston/student-life/
Annual oktoberfest and honk festival *honk! festival*	ef.com/ia/destinations/united-states/boston/student-life/
Boston jazz week	telegraph.co.uk/travel/[...]/Boston-attractions.html
Boston ballet	celebrateboston.com/events.htm
Fourth of july reading of the declaration of independence	ef.com/ia/destinations/united-states/boston/student-life/
Isabella stewart gardner museum	hotels-rates.com/hotels_reservations/property/27353/
Revere beach sand sculpting festival	bizbash.com/bostons_top_100_events/boston/story/21513/
Shakespeare on the common	boston-discovery-guide.com/boston-event-calendar-[...]
Boston bacon takedown	remarkablebostonevents.blogspot.com/[...]
Jazz at the fort	berklee.edu/events/summer/
Cambridge dance party	cheapthrillsboston.blogspot.com/[...]
Boston celtic music festival	ef.com/ia/destinations/united-states/boston/student-life/
Taste of the south end	bizbash.com/bostons_top_100_events/boston/story/21513/
Greenway open market	travel2boston.us/boston-harborfest-30th-anniversary-[...]
Boston winter jubilee	ef.com/ia/destinations/united-states/boston/student-life/
Urban ag fair	bostonmamas.com/2009/09/
Figment boston	festivaltrek.com/festivals-location/USA/Massachusetts/
Boston kite festival	bostoneventsinsider.com/2009/06/
Chefs in shorts	bizbash.com/bostons_top_100_events/boston/story/21513/
Old south meeting house	hotels-rates.com/hotels_reservations/property/27353/

Figure 3: Items and their source sites from the top of the ranked list for the Boston events experiment. Superscript numbers indicate the iteration at which the item was added to the seed via implicit feedback. "[...]" indicates the URL was truncated to fit in the figure. To improve readability, duplicate items were grouped and placed in italics.

Item	**Source**
^{0}Challah *braided challah*	(original seed)
^{3}Potato latkes *latkes; sweet potato latkes; potato latke*	jewishveg.com/recipes.html
^{1}Blintzes *cheese blintzes; blintz*	jewfaq.org/food.htm
^{0}Knishes *potato knishes; knish*	(original seed)
^{2}Noodle kugel *noodle kugel recipe; kugel; sweet noodle kugel*	pinterest.com/foodnwine/jewish-foods-holidays/
^{4}Tzimmes *carrot tzimmes*	jewfaq.org/food.htm
Matzo balls *matzo ball soup; matzo; matzoh balls*	jewishveg.com/recipes.html
Potato kugel	challahconnection.com/recipe.asp
Passover recipes *hanukkah recipes*	lynnescountrykitchen.net/jewish/index.html
Gefilte fish	jewfaq.org/food.htm
Honey cake	kveller.com/activities/food/Holidays.shtml
Soups, kugels & liver	allfreshkosher.com/freezer/blintzes-knishes-burekas.html
Charoset *haroset*	jewishveg.com/recipes.html
Hamantaschen	butterfloureggs.com/recipes/challah/
Matzo meal	glattmart.net/en/198-blintzes
Rugelach *rugelach recipe*	pinterest.com/foodnwine/jewish-foods-holidays/
Matzo brei	ilovekatzs.com/breakfast-houston/
Cholent	jewfaq.org/food.htm
Sufganiyot	kosheronabudget.com/kosher-recipe-exchange-the-complete-list/
Potato pancakes	jewishveg.com/recipes.html
Noodle pudding	epicurious.com/articlesguides/holidays/[...]/yomkippur_recipes
Kreplach	allmenus.com/md/pikesville/245371-suburban-house/menu/
Barley soup	ecampus.com/love-knishes-irrepressible-guide-jewish/[...]
Mushroom barley *mushroom barley soup*	zagat.com/r/veselka-manhattan-0/menu
Chopped liver	ryedeli.com/food/here
Garlic mashed potatoes	tovascatering.com/menu_brochure.html
Caponata	lynnescountrykitchen.net/jewish/index.html
Compote	kveller.com/activities/food/Holidays.shtml
Farfel & mushrooms *farfel*	hungariankosher.com/fb/catering-list.html
Kasha varnishkes	jinsider.com/videos/vid/496-recipes/1867-potato-knishes.html

Figure 5: Items and their source sites from the top of the ranked list for the Jewish foods experiment.

There is good deal of work in the machine learning community on aggregating ranked lists (*e.g.*, Dwork et al., 2001). These are lists that are typically already cleaned, fixed in scope, and ranked by individual experts, unlike our case. There is also a body of work on aggregated search (Beg and Ahmad, 2003; Hsu and Taksa, 2005; Lalmas, 2011), which typically uses a text query to aggregate results from multiple search engines, or of multiple formats or domains (e.g. image and news), and returns links to the full source. Our goal is not to rank URLs but to scrape out and rank information gleaned from them. There are many resources for performing a search or query by example. They often involve using a single example of a full document (Chang and Lui, 2001; Liu et al., 2003; Wang and Lochovsky, 2003; Zhai and Liu, 2005) or image (Smeulders et al., 2000), in order to retrieve more documents, structures within documents, or images. "Query by example" can also refer to methods of creating formal database queries from

Jewish foods
Knishes
Challah
Crackers
Dinner rolls
Focaccie
Pains sucres
Pains plats
Biscotti integral de algarroba
Souffle de zanahorias
Tarta de esparragos
Leftover meat casserole
Pan de canela
Focaccia
Sweet hobz
Pranzu rolls
Focacce
Chicken quesadillas
Baked chicken chimichangas
Honey mustard salad dressing
Dixxijiet hobz
Roast partridge
Fanny farmer brownies
Pan pratos
Pan doce
Cea rolls
Flat paes
Hobz dixxijiet

Figure 6: Google Sets results for the Jewish foods experiment (seed italicized).

user input text; none of these is the task we explore here.

Methods such as Gupta and Sarawagi (2009) and Pantel et al. (2009) involve growing a list, but require preprocessing which crawls the web and creates an index of HTML lists in an unsupervised manner. We do not preprocess, instead we perform information extraction online, deterministically, and virtually instantaneously given access to a search engine. There is no restriction to HTML list structures, or need for more time consuming learning methods (Freitag, 1998; Soderland et al., 1999). We also do not require human-labeled web pages like *wrapper induction* methods (Kushmerick, 1997). The works of Wang and Cohen (2007, 2008) at first appear similar to ours, but differ in many significant ways such as how the seed is used, the feature space construction, and the ranking method. We tried the method of Wang and Cohen (2007, 2008) through their Boo!Wa! interface, and found that it did not perform well on our queries.

6 Conclusions

We applied a collection of machine learning techniques to solve a real problem: growing a list using the Internet. The gold standard experiments showed that our method can perform well on a wide range of list growing problems. In our open-ended experiments, we found that the algorithm produced meaningful lists, with information extracted from a wide variety of sources, that compared favorably with lists from existing related technology. Finally, we presented a theoretical bound that justifies our use of Bayesian Sets in a setting where its feature independence assumptions are not met. The problem of aggregating expert knowledge in the form of lists on the Internet is important in many domains and our algorithm is a promising large scale solution that can be immediately implemented and used.

References

Beg, M. M. S. and Ahmad, N. (2003). Soft computing techniques for rank aggregation on the world wide web. *World Wide Web*, 6(1):5–22.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Chang, C.-H. and Lui, S.-C. (2001). Iepad: Information extraction based on pattern discovery. In *Proceedings of WWW*.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of WWW*.

Freitag, D. (1998). Information extraction from HTML: application of a general machine learning approach. In *Proceedings of AAAI*.

Ghahramani, Z. and Heller, K. A. (2005). Bayesian sets. In *Proceedings of NIPS*.

Gupta, R. and Sarawagi, S. (2009). Answering table augmentation queries from unstructured lists on the web. *Proceedings of the VLDB Endowment*.

Hsu, D. F. and Taksa, I. (2005). Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, 8(3):449–480.

Jindal, P. and Roth, D. (2011). Learning from negative examples in set-expansion. In *Proceedings of ICDM*.

Kushmerick, N. (1997). *Wrapper induction for information extraction*. PhD thesis, University of Washington.

Lalmas, M. (2011). Aggregated search. In Melucci, M. and Baeza-Yates, R., editors, *Advanced Topics on Information Retrieval*. Springer.

Liu, B., Grossman, R., and Zhai, Y. (2003). Mining data records in web pages. In *Proceedings of SIGKDD*.

Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.-M., and Vyas, V. (2009). Web-scale distributional similarity and entity set expansion. In *Proceedings of Empirical Methods in Natural Language Processing*.

Sadamitsu, K., Saito, K., Imamura, K., and Kikui, G. (2011). Entity set expansion using topic information. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Sarmento, L., Jijkoun, V., de Rijke, M., and Oliveira, E. (2007). More like these : growing entity classes from seeds. In *Proceedings of CIKM*.

Smeulders, A. W. M., Member, S., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380.

Soderland, S., Cardie, C., and Mooney, R. (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272.

Tran, M.-V., Nguyen, T.-T., Nguyen, T.-S., and Le, H.-Q. (2010). Automatic named entity set expansion using semantic rules and wrappers for unary relations. In *Proceedings of IALP*.

Wang, J. and Lochovsky, F. H. (2003). Data extraction and label assignment for web databases. In *Proceedings of WWW*.

Wang, R. C. and Cohen, W. W. (2007). Language-independent set expansion of named entities using the web. In *Proceedings of ICDM*.

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

Zhai, Y. and Liu, B. (2005). Web data extraction based on partial tree alignment. In *Proceedings of WWW*.

Zhang, L. and Liu, B. (2011). Entity set expansion in opinion documents. In *ACM Conference on Hypertext and Hypermedia*.

Supplement to Growing a List

This supplementary material expands on the algorithm, experiments, and theory given in the main text of Growing a List. In Section 1 we give implementation details for our algorithm. In Section 2 we give further detail on the Wikipedia gold standard experiments, and provide three additional sets of open-ended experiments (smartphone apps, politicians, and machine learning conferences). In Section 3 we give the proof of our main theoretical result, Theorem 1, as well as additional theoretical results, including a Hoeffding's-based generalization bound.

1 Implementation Details

Source discovery: This step requires submitting the query *"term1" "term2"* to a search engine. In our experiments we used Google as the search engine, but any index would suffice. We retrieved the top 100 results.

List extraction: For each site found with the combinatorial search, we look for HTML tags around the seed items. We use the following lines of HTML to illustrate:

```
<h2><b><a href="example1.com"> Boston Harborfest</a></b></h2>
<b><a href="example2.com"> Jimmy fund scooper bowl </a></b>
<b><a href ="example3.com"> the Boston Arts Festival 2012</a></b>
<h3><b><a href="example4.com"> Boston bacon takedown </a></b></h3>
<a href="example5.com"> Just a url </a>
```

For each of the two seed items used to discover this source, we search the HTML for the pattern:

<largest set of HTML tags>(up to 5 words) seed item (up to 5 words)<matching end tags>.

In the above example, if the first seed item is "Boston arts festival," then it matches the pattern with the HTML tags: `<b><a>`. If the second seed item is "Boston harborfest," it matches the pattern with HTML tags: `<h2><b><a>`. We then find the largest set of HTML tags that are common to both seed items, for this site. In this example, "Boston arts festival" does not have the `<h2>` tag, so the largest set of common tags is: `<b><a>`. If there are no HTML tags common to both seed items, we discard the site. Otherwise, we extract all items on the page that use the same HTML tags. In this example, we extract everything with both a `<b>` and an `<a>` tag, which means "Jimmy fund scooper bowl" and "Boston bacon takedown," but not "Just a url."

In our experiments, to avoid search spam sites with extremely long lists of unrelated keywords, we reject sources that return more than 300 items. We additionally applied a basic filter rejecting items of more than 60 characters or items consisting of only numbers and punctuation. No other processing was done.

Feature Space: We do separate Google searches for each item we have extracted to find the set of webpages containing it. We use quotes around the query term and discard results when Google's spelling correction system modifies the query. Our ranking algorithm gauges whether an item appears on a similar set of websites to the seed, so it is essential to consider the websites *without* an overlap between the item and the seed. We retrieve the top 300 search results.

Ranking: Recall the scoring function that we use to rank retrieved items by relevance:

$$f_S(x) = \frac{1}{Z(m)} \sum_{j=1}^{N} x_j \log \frac{\alpha_j + \sum_{s=1}^{m} x_j^s}{\alpha_j} + (1 - x_j) \log \frac{\beta_j + m - \sum_{s=1}^{m} x_j^s}{\beta_j}. \tag{S1}$$

As is typically the case in Bayesian analysis, there are several options for selecting the prior hyperparameters α_j and β_j, including the non-informative prior $\alpha_j = \beta_j = 1$. Heller & Ghahramani (2006) recommend using the empirical distribution. Given n items to score $x^1, \ldots, x^n$, we let

$$\alpha_j = \kappa_1 \left(\frac{1}{n} \sum_{i=1}^{n} x_j^i \right), \quad \beta_j = \kappa_2 \left(1 - \frac{1}{n} \sum_{i=1}^{n} x_j^i \right). \tag{S2}$$

The first term in the sum in (S1) corresponds to the amount of score obtained by x for the co-occurrence of feature j with the seed, and the second term corresponds to the amount of score obtained for the non-occurrence of feature j with the seed. When $\alpha_j = \beta_j$, the amount of score obtained when x_j and the seed both occur is equivalent to the amount of score obtained when x_j and the seed both do not occur. Increasing β_j relative to α_j gives higher emphasis to co-occurring features. This is useful when the feature vectors are very sparse, as they are here; thus we take $\kappa_2 > \kappa_1$. Specifically, in all of our experiments we took $\kappa_2 = 5$ and $\kappa_1 = 2$, similarly to that done in Heller & Ghahramani (2006).

Feedback: To avoid filling the seed with duplicate items like "Boston arts festival" and "The boston arts festival 2012," in our implicit feedback we do not add items to the seed if they are a sub- or super-string of a current seed item.

Our algorithm leverages the speed of Google, however, Google creates an artificial restriction on the number of queries one can make per minute. This, and the speed of our internet connection in downloading webpages, are the only two slow steps in our method - when the webpages are already downloaded, the whole process takes seconds. Both issues would be fixed if we had our own index and search engine. On the other hand, for curating master lists, the results are just as useful whether or not they are obtained instantaneously.

2 Additional Experimental Results

Here we give additional details relating to the Wikipedia gold standard experiments, and provide results for additional open-ended experiments.

2.1 Wikipedia Gold Standard Experiments

In Table 2.1 we give a completion enumeration of the results from the Wikipedia gold standard experiments. For each list growing problem, we provide the Precision@5 and the Precision@20 for all three methods (our method, Google Sets, and Boo!Wa!). This table illustrates both the diversity of the sampled list growing problems and the substantially improved performance of our method compared to the others.

2.2 Additional open-ended experiments

We present results from three additional open-ended experiments: smartphone apps, politicians, and machine learning conferences. These experiments were done with the same algorithm and parameter settings as all of the experiments in the main text; only the seed items were changed.

2.2.1 Apps

In this experiment, we began with two popular apps as the seed items: "Word lens" and "Aroundme." We ran the algorithm for 5 iterations, throughout which 7,630 items were extracted. Figure S1 shows the top 50 most highly ranked items, together with the source site where they were discovered. Not only are the results almost exclusively apps, but they come from a wide variety of sources including personal sites, review sites, blogs, and news sites. In Figure S3, we show the complete list of results returned from Google Sets for the same seed, which contains a small list of apps. Boo!Wa! was unable to return any results for this seed.

Table S1: Results for all 50 experiments with Wikipedia gold standards. “Us” indicates our method, “BW” indicates Boo!Wa!, and “GS” indicates Google Sets. “List of” has been removed from the title of each Wikipedia article, for brevity.

	Precision@5			**Precision@20**		
Wikipedia gold standard list	**Us**	**BW**	**GS**	**Us**	**BW**	**GS**
Awards and nominations received by Chris Brown	1	1	0	0.95	0.95	0
Medal of Honor recipients educated at the United States Military Academy	0.2	0	0	0.2	0.05	0
Nine Inch Nails concert tours	0.4	0	0	0.55	0	0
Bleach episodes (season 4)	0	0	0	0	0	0
Storms in the 2005 Atlantic hurricane season	0.2	0	0	0.2	0	0
Houses and associated buildings by John Douglas	0.6	0.8	0	0.55	0.7	0
Kansas Jayhawks head football coaches	1	0.8	0	1	0.95	0
Kraft Nabisco Championship champions	0.2	0	0	0.15	0	0
Washington state symbols	0	0	0	0	0	0
World Heritage Sites of the United Kingdom	0.4	0	0	0.35	0	0
Philadelphia Eagles head coaches	0	0	0	0.05	0	0
Los Angeles Dodgers first-round draft picks	0.8	0	0	0.5	0	0.05
New York Rangers head coaches	0.2	0.8	0	0.2	0.75	0
African-American Medal of Honor recipients	1	0	0	0.95	0	0
Current sovereign monarchs	0.6	0	0	0.5	0	0
Brotherhood episodes	1	0.4	0	0.65	0.3	0
Knight’s Cross of the Iron Cross with Oak Leaves recipients (1945)	0	0	0	0	0	0.05
Pittsburgh Steelers first-round draft picks	0.2	0	0	0.5	0	0
Tallest buildings in New Orleans	0.4	0	0.6	0.4	0	0.15
Asian XI ODI cricketers	0.2	0	0.4	0.1	0	0.15
East Carolina Pirates head football coaches	0.2	0.2	0	0.05	0.05	0
Former championships in WWE	0.4	0	0.4	0.35	0.05	0.3
Space telescopes	0	0	0	0	0	0
Churches preserved by the Churches Conservation Trust in Northern England	0	0	0	0	0	0
Canadian Idol finalists	0.6	0	0.2	0.65	0	0.2
Wilfrid Laurier University people	1	0	0	0.9	0	0
Wario video games	0.2	0.6	0.8	0.25	0.35	0.4
Governors of Washington	0.8	0	0	0.6	0	0
Buffalo Sabres players	0.2	0	0	0.15	0	0
Australia Twenty20 International cricketers	0.4	0	1	0.5	0	0.7
Awards and nominations received by Madonna	1	1	0.2	0.95	1	0.05
Yukon Quest competitors	0.6	0.4	0.2	0.5	0.55	0.05
Arsenal F.C. players	0.8	0	0	0.95	0	0
Victoria Cross recipients of the Royal Navy	0.2	0	0	0.25	0	0
Formula One drivers	0	0.6	1	0	0.65	0.6
Washington & Jefferson College buildings	0	0	0	0	0	0
X-Men video games	0.4	0.8	0	0.3	0.3	0
Governors of Florida	0.6	0	0	0.5	0	0
The Simpsons video games	0	0	0	0.05	0	0
Governors of New Jersey	0.8	0.2	0	0.5	0.05	0
Uncharted characters	0.8	0	0.8	0.5	0	0.65
Miami Marlins first-round draft picks	0.8	1	0	0.6	0.3	0
Tallest buildings in Dallas	0.4	0.2	0	0.45	0.05	0
Cities and towns in California	0.8	0.6	1	0.8	0.15	0.9
Olympic medalists in badminton	0.6	0	0	0.35	0	0
Delegates to the Millennium Summit	0.6	0.6	0	0.8	0.3	0
Honorary Fellows of Jesus College, Oxford	0.8	0.4	0	0.95	0.6	0
Highlander: The Raven episodes	0.2	1	0	0.1	0.9	0
Voice actors in the Grand Theft Auto series	0.2	0	0	0.2	0	0
Medal of Honor recipients for the Vietnam War	0.8	0.8	0	0.95	0.3	0

2.2.2 Politicians

In this experiment, we began with two politicians as the seed items: “Barack obama” and “Scott brown.” We ran the algorithm for 5 iterations, yielding 8,384 items. Figure S2 shows the top 50 most highly ranked items, together with the source site where they were discovered. All of the items in our list are names of politicians or politically influential individuals. In Figure S4, we show the results returned from Google Sets for the same seed, which contain only a few people related to politics. Boo!Wa! was unable to return any

Item	**Source**
^{0}Word lens	(original seed)
^{2}Read it later	iapps.scenebeta.com/noticia/ultrasn0w
read later	
^{0}Aroundme	(original seed)
^{3}Instapaper	time.com/time/specials/packages/completelist/0,29569,2044480,00.html
instapaper app	
^{4}Evernote	crosswa.lk/users/amberreyn/iphone
evernote app	
^{1}Flipboard	crosswa.lk/users/amberreyn/iphone
Dolphin browser	1mobile.com/maps-4530.html
Skitch	worldwidelearn.com/education-articles/top-50-apps-for-time-management.html
Facebook messenger	crosswa.lk/users/amberreyn/iphone
Zite	adriandavis.com/blog/bid/118125/What-s-Installed-on-My-iPad
Tweetbot	duckduckgo.com/1/c/IOS_software
Google currents	secure.crosswa.lk/users/osservatorio/iphone
Springpad	time.com/time/specials/packages/completelist/0,29569,2044480,00.html
Imessage	iphoneae.com/cydia/ihacks.html
Retina display	twicpic.blogspot.com/
Ibooks	crosswa.lk/users/amberreyn/iphone
Dropbox	mobileappreviews.craveonline.com/reviews/apple/191-word-lens
dropbox (app); dropbox app	
Marco arment	wired.com/gadgetlab/tag/instapaper/
Doubletwist	appolicious.com/finance/articles/4500-new-smartphone-[...]-download-these-apps-first
Google latitude	iapps.scenebeta.com/noticia/ultrasn0w
Gowalla	mobileappreviews.craveonline.com/reviews/apple/191-word-lens
Skype for ipad	secure.crosswa.lk/users/osservatorio/iphone
Hulu plus	appadvice.com/appnn/2010/12/expect-2011-app-store
Icloud	thetechcheck.com/tag/iphone-apps/
Qik video	1mobile.com/maps-4530.html
qik	
Find my friends	oradba.ch/2012/05/ipad-apps/
Skydrive	crosswa.lk/users/MyCsPiTTa/iphone
Google shopper	mobileappreviews.craveonline.com/reviews/apple/191-word-lens
Swype	techcrunch.com/2011/01/21/congratulations-crunchies-winners-twitter-takes-best-[...]
Pulse news reader	techcrunch.com/2011/01/21/congratulations-crunchies-winners-twitter-takes-best-[...]
Spotify	crosswa.lk/users/amberreyn/iphone
Readability	tips.flipboard.com/2011/12/08/iphone-user-guide/
Apple app store	socialmediaclub.org/blogs/from-the-clubhouse/finding-'perfect-app'-your-mobile-device
Tweetdeck	iapps.scenebeta.com/noticia/ultrasn0w
Angry birds space	appys.com/reviews/aroundme/
Smartwatch	theverge.com/2012/3/3/2839985/[...]-client-free-mac-app-store
Vlingo	mobileappreviews.craveonline.com/reviews/apple/191-word-lens
Rdio	techcrunch.com/2011/01/21/congratulations-crunchies-winners-twitter-takes-best-[...]
Google goggles	sofialys.com/newsletter_sofialys/
Xmarks	40tech.com/tag/read-it-later/
Ios 6	zomobo.net/evernote-hello
Ibooks author	duckduckgo.com/1/c/IOS_software
Google drive	geekandgirliestuff.blogspot.com/2012/01/instapaper-readitlater-readability.html
Facetime	bgpublishers.com.au/2011/10/

Figure S1: Items and their source sites from the top of the ranked list for the apps experiment.

results for this seed.

2.2.3 Machine Learning Conferences

In this experiment, we began with two machine learning conferences as the seed items: "International conference on machine learning," and "Neural information processing systems." We ran the algorithm for 5 iterations, yielding 3,791 items. Figure S5 shows the top 50 most highly ranked items, together with the source site where they were discovered. A number of popular machine learning conferences, as well as journals, are at the top of the list. Many of the sources are the sites of machine learning researchers. In Figure S6, we show the results returned from Google Sets for the same seed of two conferences. Google Sets returned a small list containing some conferences, but the list is less complete and some of the conferences are not closely related to machine learning. Boo!Wa! was unable to return any results for this seed.

3 Proofs and Additional Theoretical Results

In this section, we provide an alternative to Theorem 1 that uses Hoeffding's inequality (Theorem S1), the proof of Theorem 1, comments on the effect of the prior ($\gamma_{\min}$) on generalization, and an example showing

Item	**Source**
^{0}Barack obama	(original seed)
obama	
^{0}Scott brown	(original seed)
^{1}John kerry	publicpolicypolling.com/main/scott-brown/
^{3}Barney frank	masslive.com/politics/index.ssf/2012/03/sens_scott_brown_and_john_kerr.html
^{4}John mccain	publicpolicypolling.com/main/scott-brown/
mccain	
^{2}Nancy pelosi	theladypatriot.com/
pelosi	
Mitch mcconnell	publicpolicypolling.com/main/scott-brown/
Joe lieberman	publicpolicypolling.com/main/scott-brown/
Mike huckabee	publicpolicypolling.com/main/scott-brown/
Mitt romney	masslive.com/politics/index.ssf/2012/04/power_of_incumbency_boasts_sen.html
Bill clinton	mediaite.com/online/nothing-but-net-sen-scott-brown-makes-a-half-court-shot-at-local-community-[...]
John boehner	audio.wrko.com/a/50487720/why-did-scott-brown-agree-with-barack-obama-s-recess-appointment.htm
boehner	
Hillary clinton	blogs.wsj.com/washwire/2010/01/29/all-in-the-family-obama-related-to-scott-brown/
Jon kyl	tpmdc.talkingpointsmemo.com/nancy-pelosi/2010/08/
Joe biden	publicpolicypolling.com/main/scott-brown/
Rudy giuliani	publicpolicypolling.com/main/scott-brown/
Harry reid	theladypatriot.com/
Olympia snowe	publicpolicypolling.com/main/scott-brown/
Lindsey graham	politico.com/news/stories/0410/36112.html
Newt gingrich	masspoliticsprofs.com/tag/barack-obama/
Jim demint	theladypatriot.com/
Arlen specter	theladypatriot.com/
Dick cheney	blogs.wsj.com/washwire/2010/01/29/all-in-the-family-obama-related-to-scott-brown/
George w bush	wellgroomedmanscape.com/tag/scott-brown/
george w. bush	
Eric holder	disruptthenarrative.com/category/john-kerry/
Dennis kucinich	publicpolicypolling.com/main/scott-brown/
Timothy geithner	tpmdc.talkingpointsmemo.com/john-mccain/
Barbara boxer	publicpolicypolling.com/main/scott-brown/
Tom coburn	itmakessenseblog.com/tag/nancy-pelosi/
Orrin hatch	publicpolicypolling.com/main/scott-brown/
Michael bloomberg	masspoliticsprofs.com/tag/barack-obama/
Elena kagan	audio.wrko.com/a/50487720/why-did-scott-brown-agree-with-barack-obama-s-recess-appointment.htm
Maxine waters	polination.wordpress.com/category/nancy-pelosi/
Al sharpton	porkbarrel.tv/
Rick santorum	audio.wrko.com/a/50487720/why-did-scott-brown-agree-with-barack-obama-s-recess-appointment.htm
Ted kennedy	newomenforchange.org/tag/scott-brown/
Janet napolitano	disruptthenarrative.com/category/john-kerry/
Jeff sessions	tpmdc.talkingpointsmemo.com/john-mccain/
Jon huntsman	publicpolicypolling.com/main/scott-brown/
Michele bachmann	publicpolicypolling.com/main/scott-brown/
Al gore	publicpolicypolling.com/main/scott-brown/
Rick perry	publicpolicypolling.com/main/scott-brown/
Eric cantor	publicpolicypolling.com/main/scott-brown/
Ben nelson	publicpolicypolling.com/main/scott-brown/
Karl rove	politico.com/news/stories/1010/43644.html

Figure S2: Items and their source sites from the top of the ranked list for the politicians experiment.

that Bayesian Sets does not satisfy the requirements for "uniform stability" defined by Bousquet & Elisseeff (2002).

3.1 An Alternate Generalization Bound

We begin by showing that the normalized score $f_S(x)$ in (S1) takes values only on $[0, 1]$.

Lemma S1. *$0 \leq f_S(x) \leq 1$.*

Apps
Word lens
Aroundme
Lifestyle
View in itunes
Itunes
Jcpenney weekly deals
Coolibah digital scrapbooking
Epicurious recipes shopping list
170000 recipes bigoven
Cf iviewer
Txtcrypt
Speak4it
Off remote free
Catholic calendar
Gucci
Board
Ziprealty real estate
Allsaints spitalfields
Lancome make up
Pottery barn catalog viewer
Amazon mobile
Gravity clock
Dace
Zara
Style com
Iridiumhd
Ebanner lite
Mymemoir
Rezepte
Maxjournal for ipad
Chakra tuning
My secret diary
Pretty planner
Remodelista
Ipause

Figure S3: Google Sets results for the apps experiment (seed italicized).

Politicians
Barack obama
Scott brown
Our picks movies
Sex
Department of justice
Viral video
Africa
One persons trash
Donald trump
New mom confessions
Nonfiction
Libya
Sarah palin
Mtv
Alan greenspan
Great recession
Life stories
Jon hamm
Islam
The killing
American idol
Middle east
Celebrity
Tea parties
Budget showdown

Figure S4: Google Sets results for the politicians experiment (seed italicized).

Proof. It is easy to see that $f_S(x) \geq 0$. To see that $f_S(x) \leq 1$,

$$\begin{aligned}
\max_{S,x} f_S(x) &= \frac{1}{Z(m)} \max_{S,x} \sum_{j=1}^{N} x_j \log \frac{\alpha_j + \sum_{s=1}^{m} x_j^s}{\alpha_j} + (1 - x_j) \log \frac{\beta_j + m - \sum_{s=1}^{m} x_j^s}{\beta_j} \\
&\leq \frac{1}{Z(m)} \sum_{j=1}^{N} \max_{x_j, x_j^1, \ldots, x_j^m} x_j \log \frac{\alpha_j + \sum_{s=1}^{m} x_j^s}{\alpha_j} + (1 - x_j) \log \frac{\beta_j + m - \sum_{s=1}^{m} x_j^s}{\beta_j} \\
&= \frac{1}{Z(m)} \sum_{j=1}^{N} \max \left\{ \max_{x_j^1, \ldots, x_j^m} \log \frac{\alpha_j + \sum_{s=1}^{m} x_j^s}{\alpha_j}, \max_{x_j^1, \ldots, x_j^m} \log \frac{\beta_j + m - \sum_{s=1}^{m} x_j^s}{\beta_j} \right\} \\
&= \frac{1}{Z(m)} \sum_{j=1}^{N} \max \left\{ \log \frac{\alpha_j + m}{\alpha_j}, \log \frac{\beta_j + m}{\beta_j} \right\} \\
&= \frac{1}{Z(m)} \sum_{j=1}^{N} \log \frac{\min\{\alpha_j, \beta_j\} + m}{\min\{\alpha_j, \beta_j\}} \\
&\leq \frac{1}{Z(m)} \sum_{j=1}^{N} \log \frac{\gamma_{\min} + m}{\gamma_{\min}} \\
&= 1.
\end{aligned}$$

Item	Source
^{3}Machine learning journal	cs.columbia.edu/∼rocco/papers/papers.html
Machine learning	
^{0}Neural information processing systems	(original seed)
^{4}*advances in neural information processing*	
advances in neural information processing systems	
advances in neural information	
advances in neural information processing systems (nips)	
nips (2007); nips (2008); nips 2007; nips 2009	
nips, 2007; nips, 2008	
neural information processing systems (nips 2007)	
^{0}International conference on machine learning	(original seed)
icml 2005; icml 2006; icml (2010); icml, 2010	
international conference on machine learning (icml), 2009	
international conference on machine learning, icml (2005)	
icml 2010; icml-08	
international conference on machine learning (icml), 2008	
international conference on machine learning (icml) (2008)	
27th international conference on machine learning	
^{1}European conference on machine learning	userpage.fu-berlin.de/mtoussai/publications/index.html
^{2}*conference on machine learning (ecml)*	
Journal of machine learning research	cs.princeton.edu/∼blei/publications.html
journal of machine learning research (jmlr)	
machine learning research	
Artificial intelligence and statistics	cs.princeton.edu/∼blei/publications.html
international conference on artificial intelligence and statistics	
Conference on learning theory	cs.cmu.edu/∼lafferty/publications.html
Journal of artificial intelligence research	web.eecs.umich.edu/∼baveja/rlpubs.html
Conference on uncertainty in artificial intelligence (uai)	cs.duke.edu/∼johns/
uncertainty in artificial intelligence (uai)	
conference on uncertainty in artificial intelligence	
uncertainty in artificial intelligence [uai]	
uncertainty in artificial intelligence	
Computer vision and pattern recognition	cs.princeton.edu/∼blei/publications.html
Ieee international conference on data mining (icdm)	cis.upenn.edu/∼ungar/Datamining/publications.html
Learning in graphical models	cseweb.ucsd.edu/∼saul/papers.html
Aaai (2006)	research.google.com/pubs/ArtificialIntelligence[...]
national conference on artificial intelligence	
Proceedings of the sixth acm sigkdd international conference on knowledge discovery & data mining	web.engr.oregonstate.edu/∼tgd/publications/index.html
Machine learning summer school	arnetminer.org/page/conference-rank/html/ML[...]
International joint conference on artificial intelligence	userpage.fu-berlin.de/mtoussai/publications/index.html

Figure S5: Items and their source sites from the top of the ranked list for the machine learning conferences experiment.

Machine learning conferences
International conference on machine learning
Neural information processing systems
Society for neuroscience
Vision sciences society
Optical society of america
Japan neuroscience society
Computationalneuroscience organization
Japan neural network society
Institute of image information television engineers
Vision society of japan
American association for artificial intelligence
Psychonomic society
Association for psychological science
Decision hyperplane
San mateo
Computational and systems neuroscience
International conference on automated planning and scheduling
Uncertainty in artificial intelligence
International joint conference on artificial intelligence

Figure S6: Google Sets results for the machine learning conferences experiment (seed italicized).

□

Now we provide the alternative to Theorem 1 that uses Hoeffding's inequality.

Theorem S1. *With probability at least $1-\delta$ on the draw of the training set S,*

$$\mathbb{E}_x\left[f_S(x)\right] \geq \frac{1}{m}\sum_{s=1}^{m} f_S(x^s) - \sqrt{\frac{1}{2m}\log\left(\frac{2N}{\delta}\right)}.$$

Proof. For convenience, denote the seed sample average as $\mu_j := \frac{1}{m}\sum_{s=1}^{m} x_j^s$, and the probability that $x_j = 1$ as $p_j := \mathbb{E}_x[x_j]$. Then,

$$\begin{aligned}
&\frac{1}{m}\sum_{s=1}^{m} f_S(x^s) - \mathbb{E}_x\left[f_S(x)\right] \\
&\qquad = \frac{1}{N\log\left(\frac{\gamma_{\min}+m}{\gamma_{\min}}\right)}\sum_{j=1}^{N}(\mu_j - p_j)\log\frac{\alpha_j + m\mu_j}{\alpha_j} + (p_j-\mu_j)\log\frac{\beta_j + m(1-\mu_j)}{\beta_j} \\
&\qquad \leq \frac{1}{N}\sum_{j=1}^{N}|\mu_j - p_j|. \qquad \text{(S3)}
\end{aligned}$$

For any particular feature j, Hoeffding's inequality (Hoeffding, 1963) bounds the difference between the empirical average and the expected value:

$$\mathbb{P}(|\mu_j - p_j| > \epsilon) \leq 2\exp\left(-2m\epsilon^2\right). \qquad \text{(S4)}$$

We then apply the union bound to bound the average over features:

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{N}\sum_{j=1}^{N}|\mu_j - p_j| > \epsilon\right) &\leq \mathbb{P}\left(\bigcup_{j=1}^{N}\{|\mu_j - p_j| > \epsilon\}\right) \\
&\leq \sum_{j=1}^{N}\mathbb{P}\left(|\mu_j - p_j| > \epsilon\right) \\
&\leq 2N\exp\left(-2m\epsilon^2\right). \qquad \text{(S5)}
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\frac{1}{m}\sum_{s=1}^{m} f_S(x^s) - \mathbb{E}_x\left[f_S(x)\right] > \epsilon\right) \leq 2N\exp\left(-2m\epsilon^2\right), \qquad \text{(S6)}$$

and the theorem follows directly. □

The bound in Theorem S1 has a tighter dependence on δ than the bound in Theorem 1, however it depends inversely on N, the number of features. We prefer the bound in Theorem 1, which is independent of N.

3.2 Proof of the Main Theoretical Result

We now present the proof of Theorem 1. The result uses the algorithmic stability bounds of Bousquet & Elisseeff (2002), specifically the bound for pointwise hypothesis stability. We begin by defining an appropriate loss function. Suppose x and S were drawn from the same distribution $\mathcal{D}$. Then, we wish for $f_S(x)$ to be as large as possible. Because $f_S(x) \in [0,1]$, an appropriate metric for the loss in using f_S to score x is:

$$\ell(f_S, x) = 1 - f_S(x). \qquad \text{(S7)}$$

Further, $\ell(f_S, x) \in [0,1]$.

For algorithmic stability analysis, we will consider how the algorithm's performance changes when an element is removed from the training set. We define a modified training set in which the i'th element has been removed: $S^{\setminus i} := \{x^1, \dots, x^{i-1}, x^{i+1}, \dots, x^m\}$. We then define the score of x according to the modified training set:

$$f_{S^{\setminus i}}(x) = \frac{1}{Z(m-1)} \sum_{j=1}^{N} x_j \log \frac{\alpha_j + \sum_{s \neq i} x_j^s}{\alpha_j} + (1 - x_j) \log \frac{\beta_j + (m-1) - \sum_{s \neq i} x_j^s}{\beta_j}, \tag{S8}$$

where

$$Z(m-1) = N \log \left(\frac{\gamma_{\min} + m - 1}{\gamma_{\min}} \right). \tag{S9}$$

We further define the loss using the modified training set:

$$\ell(f_{S^{\setminus i}}, x) = 1 - f_{S^{\setminus i}}(x). \tag{S10}$$

The general idea of algorithmic stability is that if the results of an algorithm do not depend too heavily on any one element of the training set, the algorithm will be able to generalize. One way to quantify the dependence of an algorithm on the training set is to examine how the results change when the training set is perturbed, for example by removing an element from the training set. The following definition of pointwise hypothesis stability, taken from Bousquet & Elisseeff (2002), states that an algorithm has pointwise hypothesis stability if, on expectation, the results of the algorithm do not change too much when an element of the training set is removed.

Definition S1 (Bousquet & Elisseeff, 2002)**.** *An algorithm has pointwise hypothesis stability η with respect to the loss function ℓ if the following holds*

$$\forall i \in \{1, \dots, m\}, \quad \mathbb{E}_S \left[|\ell(f_S, x^i) - \ell(f_{S^{\setminus i}}, x^i)| \right] \leq \eta. \tag{S11}$$

The algorithm is said to be stable if η scales with $\frac{1}{m}$.

In our theorem, we suppose that all of the data belong to the same class of "relevant" items. The framework of Bousquet & Elisseeff (2002) can easily be adapted to the single-class setting, for example by framing it as a regression problem where all of the data points have the identical "true" output value 1. The following theorem comes from Bousquet & Elisseeff (2002), with the notation adapted to our setting.

Theorem S2 (Bousquet & Elisseeff, 2002)**.** *If an algorithm has pointwise hypothesis stability η with respect to a loss function ℓ such that $0 \leq \ell(\cdot, \cdot) \leq 1$, we have with probability at least $1 - \delta$,*

$$\mathbb{E}_x \left[\ell(f_S, x) \right] \leq \frac{1}{m} \sum_{i=1}^{m} \ell(f_S, x^i) + \sqrt{\frac{1 + 12 m \eta}{2 m \delta}}. \tag{S12}$$

We now show that Bayesian Sets satisfies the conditions of Definition S1, and determine the corresponding η. The proof of Theorem 1 comes from inserting our findings for η into Theorem S2. We begin with a lemma providing a bound on the central moments of a Binomial random variable.

Lemma S2. *Let $t \sim$ Binomial(m,p) and let $\mu_k = \mathbb{E}\left[(t - \mathbb{E}[t])^k\right]$ be the k^{th} central moment. For integer $k \geq 1$, μ_{2k} and μ_{2k+1} are $O\left(m^k\right)$.*

Proof. We will use induction. For $k = 1$, the central moments are well known (*e.g.*, Johnson et al., 2005): $\mu_2 = mp(1-p)$ and $\mu_3 = mp(1-p)(1-2p)$, which are both $O(m)$. We rely on the following recursion formula (Johnson et al., 2005; Romanovsky, 1923):

$$\mu_{s+1} = p(1-p) \left(\frac{d\mu_s}{dp} + m s \mu_{s-1} \right). \tag{S13}$$

Because μ_2 and μ_3 are polynomials in p, their derivatives will also be polynomials in p. This recursion makes it clear that for all s, μ_s is a polynomial in p whose coefficients include terms involving m.

For the inductive step, suppose that the result holds for $k = s$. That is, μ_{2s} and μ_{2s+1} are $O(m^s)$. Then, by (S13),

$$\mu_{2(s+1)} = p(1-p)\left(\frac{d\mu_{2s+1}}{dp} + (2s+1)m\mu_{2s}\right). \tag{S14}$$

Differentiating μ_{2s+1} with respect to p yields a term that is $O(m^s)$. The term $(2s+1)m\mu_{2s}$ is $O(m^{s+1})$, and thus $\mu_{2(s+1)}$ is $O(m^{s+1})$. Also,

$$\mu_{2(s+1)+1} = p(1-p)\left(\frac{d\mu_{2(s+1)}}{dp} + 2(s+1)m\mu_{2s+1}\right). \tag{S15}$$

Here $\frac{d\mu_{2(s+1)}}{dp}$ is $O(m^{s+1})$ and $2(s+1)m\mu_{2s+1}$ is $O(m^{s+1})$, and thus $\mu_{2(s+1)+1}$ is $O(m^{s+1})$.

This shows that if the result holds for $k = s$ then it must also hold for $k = s+1$ which completes the proof. □

The next lemma provides a stable, $O\left(\frac{1}{m}\right)$, bound on the expected value of an important function of a binomial random variable.

Lemma S3. *For $t \sim$ Binomial(m, p) and $\alpha > 0$,*

$$\mathbb{E}\left[\frac{1}{\alpha + t}\right] = \frac{1}{\alpha + mp} + O\left(\frac{1}{m^2}\right). \tag{S16}$$

Proof. We expand $\frac{1}{\alpha+t}$ at $t = mp$:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\alpha + t}\right] &= \mathbb{E}\left[\sum_{i=0}^{\infty}(-1)^i \frac{(t-mp)^i}{(\alpha+mp)^{i+1}}\right] \\
&= \sum_{i=0}^{\infty}(-1)^i \frac{\mathbb{E}\left[(t-mp)^i\right]}{(\alpha+mp)^{i+1}} \\
&= \frac{1}{\alpha+mp} + \sum_{i=2}^{\infty}(-1)^i \frac{\mu_i}{(\alpha+mp)^{i+1}} \qquad \text{(S17)}
\end{aligned}$$

where μ_i is the i^{th} central moment and we recognize that $\mu_1 = 0$. By Lemma S2,

$$\frac{\mu_i}{(\alpha+mp)^{i+1}} = \frac{O\left(m^{\lfloor \frac{i}{2} \rfloor}\right)}{O\left(m^{i+1}\right)} = O\left(m^{\lfloor \frac{i}{2} \rfloor - i - 1}\right). \tag{S18}$$

The alternating sum in (S17) can be split into two sums:

$$\sum_{i=2}^{\infty}(-1)^i \frac{\mu_i}{(\alpha+mp)^{i+1}} = \sum_{i=2}^{\infty} O\left(m^{\lfloor \frac{i}{2} \rfloor - i - 1}\right) = \sum_{i=2}^{\infty} O\left(\frac{1}{m^i}\right) + \sum_{i=3}^{\infty} O\left(\frac{1}{m^i}\right). \tag{S19}$$

These are, for m large enough, bounded by a geometric series that converges to $O\left(\frac{1}{m^2}\right)$. □

The following three lemmas provide results that will be useful for proving the main lemma, Lemma S7.

Lemma S4. *For all $\alpha > 0$,*

$$g(\alpha, m) := \frac{\log\left(\frac{\alpha+m}{\alpha}\right)}{\log\left(\frac{\alpha+m-1}{\alpha}\right)} \tag{S20}$$

is monotonically non-decreasing in α for any fixed $m \geq 2$.

Proof. Define $a = \frac{m-1}{\alpha}$ and $b = \frac{m}{m-1}$. Observe that $a \geq 0$ and $b \geq 1$, and that for fixed m, a is inversely proportional to α. We reparameterize (S20) to

$$g(a,b) := \frac{\log(ab+1)}{\log(a+1)}. \tag{S21}$$

To prove the lemma, it is sufficient to show that $g(a,b)$ is monotonically non-increasing in a for any fixed $b \geq 1$. Well,

$$\frac{\partial g(a,b)}{\partial a} = \frac{\frac{b}{ab+1}\log(a+1) - \frac{1}{a+1}\log(ab+1)}{(\log(a+1))^2},$$

so $\frac{\partial g(a,b)}{\partial a} \leq 0$ if and only if

$$h(a,b) := (ab+1)\log(ab+1) - b(a+1)\log(a+1) \geq 0. \tag{S22}$$

$h(a,1) = (a+1)\log(a+1) - (a+1)\log(a+1) = 0$, and,

$$\begin{aligned}\frac{\partial h(a,b)}{\partial b} &= a\log(ab+1) + a - (a+1)\log(a+1)\\ &= a\left(\log(ab+1) - \log(a+1)\right) + (a - \log(a+1))\\ &\geq 0 \quad \forall a \geq 0,\end{aligned}$$

because $b \geq 1$ and $a \geq \log(1+a)\ \forall a \geq 0$. This shows that (S22) holds $\forall a \geq 0, b \geq 1$, which proves the lemma. □

Lemma S5. *For any $m \geq 2$, $t \in [0, m-1]$, $\alpha > 0$, and $\gamma_{\min} \in (0, \alpha]$,*

$$\frac{1}{Z(m)}\log\frac{\alpha+t+1}{\alpha} \geq \frac{1}{Z(m-1)}\log\frac{\alpha+t}{\alpha}. \tag{S23}$$

Proof. Denote,

$$g(t; m, \alpha) := \frac{1}{Z(m)}\log\frac{\alpha+t+1}{\alpha} - \frac{1}{Z(m-1)}\log\frac{\alpha+t}{\alpha}. \tag{S24}$$

By Lemma S4 and $\gamma_{\min} \leq \alpha$, for any $\alpha > 0$ and for any $m \geq 2$,

$$\frac{\log\left(\frac{\alpha+m}{\alpha}\right)}{\log\left(\frac{\alpha+m-1}{\alpha}\right)} \geq \frac{\log\left(\frac{\gamma_{\min}+m}{\gamma_{\min}}\right)}{\log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)} = \frac{Z(m)}{Z(m-1)}.$$

Thus,

$$\frac{\log\left(\frac{\alpha+m}{\alpha}\right)}{Z(m)} \geq \frac{\log\left(\frac{\alpha+m-1}{\alpha}\right)}{Z(m-1)}, \tag{S25}$$

which shows

$$g(m-1; m, \alpha) = \frac{1}{Z(m)}\log\frac{\alpha+m}{\alpha} - \frac{1}{Z(m-1)}\log\frac{\alpha+m-1}{\alpha} \geq 0. \tag{S26}$$

Furthermore, because $Z(m) > Z(m-1)$,

$$\frac{\partial g(t; m, \alpha)}{\partial t} = \frac{1}{Z(m)}\frac{1}{\alpha+t+1} - \frac{1}{Z(m-1)}\frac{1}{\alpha+t} < 0, \tag{S27}$$

for all $t \geq 0$. Equations S26 and S27 together show that $g(t; m, \alpha) \geq 0$ for all $t \in [0, m-1], m \geq 2$, proving the lemma. □

Lemma S6. *For any $m \geq 2$, $t \in [0, m-1]$, $\beta > 0$, and $\gamma_{\min} \in (0, \beta]$,*

$$\frac{1}{Z(m)}\log\frac{\beta+m-t}{\beta} \geq \frac{1}{Z(m-1)}\log\frac{\beta+m-1-t}{\beta}. \tag{S28}$$

Proof. Let $\tilde{t} = m - t - 1$. Then, $\tilde{t} \in [0, m-1]$ and by Lemma S5, replacing α with β,

$$\frac{1}{Z(m)} \log \frac{\beta + \tilde{t} + 1}{\beta} \geq \frac{1}{Z(m-1)} \log \frac{\beta + \tilde{t}}{\beta}. \tag{S29}$$

□

The next lemma is the key lemma that shows Bayesian Sets satisfies pointwise hypothesis stability, allowing us to apply Theorem S2.

Lemma S7. *The Bayesian Sets algorithm satisfies the conditions for pointwise hypothesis stability with*

$$\eta = \frac{1}{\log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)(\gamma_{\min} + (m-1)p_{\min})} + O\left(\frac{1}{m^2 \log m}\right). \tag{S30}$$

Proof.

$$\begin{aligned}
&\mathbb{E}_S|\ell(f_S, x^i) - \ell(f_{S^{\setminus i}}, x^i)| \\
&\quad = \mathbb{E}_S \left| f_{S^{\setminus i}}(x^i) - f_S(x^i) \right| \\
&\quad = \mathbb{E}_S \left| \frac{1}{Z(m-1)} \sum_{j=1}^{N} \left[x_j^i \log \frac{\alpha_j + \sum_{s \neq i} x_j^s}{\alpha_j} + (1 - x_j^i) \log \frac{\beta_j + (m-1) - \sum_{s \neq i} x_j^s}{\beta_j} \right] \right. \\
&\qquad \left. - \frac{1}{Z(m)} \sum_{j=1}^{N} \left[x_j^i \log \frac{\alpha_j + \sum_{s=1}^m x_j^s}{\alpha_j} + (1 - x_j^i) \log \frac{\beta_j + m - \sum_{s=1}^m x_j^s}{\beta_j} \right] \right| \\
&\quad \leq \mathbb{E}_S \sum_{j=1}^{N} x_j^i \left| \frac{1}{Z(m-1)} \log \frac{\alpha_j + \sum_{s \neq i} x_j^s}{\alpha_j} - \frac{1}{Z(m)} \log \frac{\alpha_j + \sum_{s=1}^m x_j^s}{\alpha_j} \right| \\
&\qquad + (1 - x_j^i) \left| \frac{1}{Z(m-1)} \log \frac{\beta_j + (m-1) - \sum_{s \neq i} x_j^s}{\beta_j} - \frac{1}{Z(m)} \log \frac{\beta_j + m - \sum_{s=1}^m x_j^s}{\beta_j} \right|
\end{aligned} \tag{S31}$$

$$:= \mathbb{E}_S \sum_{j=1}^{N} x_j^i \text{term}_j^1 + (1 - x_j^i) \text{term}_j^2 \tag{S32}$$

$$\begin{aligned}
&= \sum_{j=1}^{N} \mathbb{E}_{x_j^1, \ldots, x_j^m} \left[x_j^i \text{term}_j^1 + (1 - x_j^i) \text{term}_j^2 \right] \\
&= \sum_{j=1}^{N} \mathbb{E}_{x_j^{s \neq i}} \left[\text{term}_j^1 | x_j^i = 1 \right] \mathbb{P}\left(x_j^i = 1 \right) + \mathbb{E}_{x_j^{s \neq i}} \left[\text{term}_j^2 | x_j^i = 0 \right] \mathbb{P}\left(x_j^i = 0 \right) \\
&\leq \sum_{j=1}^{N} \max \left\{ \mathbb{E}_{x_j^{s \neq i}} \left[\text{term}_j^1 | x_j^i = 1 \right], \mathbb{E}_{x_j^{s \neq i}} \left[\text{term}_j^2 | x_j^i = 0 \right] \right\},
\end{aligned} \tag{S33}$$

where (S31) uses the triangle inequality, and in (S32) we define term_j^1 and term_j^2 for notational convenience. Now consider each term in (S33) separately,

$$\begin{aligned}
\mathbb{E}_{x_j^{s \neq i}} \left[\text{term}_j^1 | x_j^i = 1 \right] &= \mathbb{E}_{x_j^{s \neq i}} \left| \frac{1}{Z(m-1)} \log \frac{\alpha_j + \sum_{s \neq i} x_j^s}{\alpha_j} - \frac{1}{Z(m)} \log \frac{\alpha_j + \sum_{s \neq i} x_j^s + 1}{\alpha_j} \right| \\
&= \mathbb{E}_{x_j^{s \neq i}} \left[\frac{1}{Z(m)} \log \frac{\alpha_j + \sum_{s \neq i} x_j^s + 1}{\alpha_j} - \frac{1}{Z(m-1)} \log \frac{\alpha_j + \sum_{s \neq i} x_j^s}{\alpha_j} \right],
\end{aligned} \tag{S34}$$

where we have shown in Lemma S5 that this quantity is non-negative. Because $\{x^s\}$ are independent, $\{x_j^s\}$ are independent for fixed j. We can consider $\{x_j^s\}_{s \neq i}$ to be a collection of $m-1$ independent Bernoulli random variables with probability of success $p_j = \mathbb{P}_{x\sim\mathcal{D}}(x_j = 1)$, the marginal distribution. Let $t = \sum_{s\neq i} x_j^s$, then $t \sim \text{Binomial}(m-1, p_j)$. Continuing (S34),

$$\begin{aligned}
\mathbb{E}_{x_j^{s\neq i}}\left[\text{term}_j^1 | x_j^i = 1\right] &= \mathbb{E}_{t\sim\text{Bin}(m-1,p_j)}\left[\frac{1}{Z(m)}\log\frac{\alpha_j+t+1}{\alpha_j} - \frac{1}{Z(m-1)}\log\frac{\alpha_j+t}{\alpha_j}\right] \\
&\leq \frac{1}{Z(m-1)}\mathbb{E}_{t\sim\text{Bin}(m-1,p_j)}\left[\log\frac{\alpha_j+t+1}{\alpha_j+t}\right] \\
&= \frac{1}{Z(m-1)}\mathbb{E}_{t\sim\text{Bin}(m-1,p_j)}\left[\log\left(1+\frac{1}{\alpha_j+t}\right)\right] \\
&\leq \frac{1}{Z(m-1)}\log\left(1+\mathbb{E}_{t\sim\text{Bin}(m-1,p_j)}\left[\frac{1}{\alpha_j+t}\right]\right) \\
&= \frac{1}{Z(m-1)}\log\left(1+\frac{1}{\alpha_j+(m-1)p_j}+O\left(\frac{1}{m^2}\right)\right). \qquad \text{(S35)}
\end{aligned}$$

The second line uses $Z(m) \geq Z(m-1)$, the fourth line uses Jensen's inequality, and the fifth line uses Lemma S3. Now we turn to the other term.

$$\begin{aligned}
&\mathbb{E}_{x_j^{s\neq i}}\left[\text{term}_j^2 | x_j^i = 0\right] \\
&\quad = \mathbb{E}_{x_j^{s\neq i}}\left|\frac{1}{Z(m-1)}\log\frac{\beta_j+(m-1)-\sum_{s\neq i}x_j^s}{\beta_j} - \frac{1}{Z(m)}\log\frac{\beta_j+m-\sum_{s\neq i}x_j^s}{\beta_j}\right| \\
&\quad = \mathbb{E}_{x_j^{s\neq i}}\left[\frac{1}{Z(m)}\log\frac{\beta_j+m-\sum_{s\neq i}x_j^s}{\beta_j} - \frac{1}{Z(m-1)}\log\frac{\beta_j+(m-1)-\sum_{s\neq i}x_j^s}{\beta_j}\right]. \qquad \text{(S36)}
\end{aligned}$$

We have shown in Lemma S6 that this quantity is non-negative. Let $q_j = 1 - p_j$. Let $t = m - 1 - \sum_{s\neq i} x_j^s$, then $t \sim \text{Binomial}(m-1, q_j)$. Continuing (S36):

$$\begin{aligned}
\mathbb{E}_{x_j^{s\neq i}}\left[\text{term}_j^2 | x_j^i = 0\right] &\leq \frac{1}{Z(m-1)}\mathbb{E}_{t\sim\text{Bin}(m-1,q_j)}\left[\log\frac{\beta_j+t+1}{\beta_j+t}\right] \\
&\leq \frac{1}{Z(m-1)}\log\left(1+\frac{1}{\beta_j+(m-1)q_j}+O\left(\frac{1}{m^2}\right)\right). \qquad \text{(S37)}
\end{aligned}$$

where the steps are as with (S35). We now take (S35) and (S37) and use them to continue (S33):

$$\begin{aligned}
&\mathbb{E}_S|\ell(f_S, x^i) - \ell(f_{S^{\setminus i}}, x^i)| \\
&\quad \leq \sum_{j=1}^N \max\left\{\frac{1}{Z(m-1)}\log\left(1+\frac{1}{\alpha_j+(m-1)p_j}+O\left(\frac{1}{m^2}\right)\right),\right. \\
&\qquad\qquad \left.\frac{1}{Z(m-1)}\log\left(1+\frac{1}{\beta_j+(m-1)q_j}+O\left(\frac{1}{m^2}\right)\right)\right\} \\
&\quad \leq \sum_{j=1}^N \frac{1}{Z(m-1)}\log\left(1+\frac{1}{\min\{\alpha_j,\beta_j\}+(m-1)\min\{p_j,q_j\}}+O\left(\frac{1}{m^2}\right)\right) \\
&\quad \leq \frac{N}{Z(m-1)}\log\left(1+\frac{1}{\gamma_{\min}+(m-1)p_{\min}}+O\left(\frac{1}{m^2}\right)\right) \\
&\quad := \eta. \qquad \text{(S38)}
\end{aligned}$$

Using the Taylor expansion of $\log(1+x)$,

$$\begin{aligned}
\eta &= \frac{N}{Z(m-1)}\left(\frac{1}{\gamma_{\min}+(m-1)p_{\min}}+O\left(\frac{1}{m^2}\right)-\frac{1}{2}\left(\frac{1}{\gamma_{\min}+(m-1)p_{\min}}+O\left(\frac{1}{m^2}\right)\right)^2\right)\\
&= \frac{N}{Z(m-1)}\left(\frac{1}{\gamma_{\min}+(m-1)p_{\min}}+O\left(\frac{1}{m^2}\right)\right)\\
&= \frac{1}{\log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)(\gamma_{\min}+(m-1)p_{\min})}+O\left(\frac{1}{m^2\log m}\right). \qquad \text{(S39)}
\end{aligned}$$

□

The proof of Theorem 1 is now a straightforward application of Theorem S2 using the result of Lemma S7.

Proof of Theorem 1. By Lemma S7, we can apply Theorem S2 to see that with probability at least $1-\delta$ on the draw of S,

$$\begin{aligned}
\mathbb{E}_x\left[\ell(f_S,x)\right] &\leq \frac{1}{m}\sum_{i=1}^{m}\ell(f_S,x^i)+\sqrt{\frac{1+12m\eta}{2m\delta}}\\
\mathbb{E}_x\left[1-f_S(x)\right] &\leq \frac{1}{m}\sum_{s=1}^{m}(1-f_S(x^s))+\sqrt{\frac{1+12m\eta}{2m\delta}}\\
\mathbb{E}_x\left[f_S(x)\right] &\geq \frac{1}{m}\sum_{s=1}^{m}f_S(x^s)-\sqrt{\frac{1+12m\eta}{2m\delta}}\\
&= \frac{1}{m}\sum_{s=1}^{m}f_S(x^s)\\
&\quad -\sqrt{\frac{1}{2m\delta}+\frac{6}{\delta\log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)(\gamma_{\min}+(m-1)p_{\min})}+O\left(\frac{1}{\delta m^2\log m}\right)}.
\end{aligned}$$

□

3.3 Comments on the effect of the prior on generalization.

The prior influences the generalization bound via the quantity

$$h(\gamma_{\min},m,p_{\min}) := \log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)(\gamma_{\min}+(m-1)p_{\min}). \qquad \text{(S40)}$$

As this quantity increases, the bound becomes tighter. We can thus study the influence of the prior on generalization by studying the behavior of this quantity as $\gamma_{\min}$ varies. The second term, $(\gamma_{\min}+(m-1)p_{\min})$, is similar to many results from Bayesian analysis in which the prior plays the same role as additional data. This term is *increasing* with $\gamma_{\min}$, meaning it yields a tighter bound with a stronger prior. The first term, $\log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)$, is inherited from the normalization $Z(m)$. This term is *decreasing* with $\gamma_{\min}$, that is, it gives a tighter bound with a weaker prior. The overall effect of $\gamma_{\min}$ on generalization depends on how these two terms balance each other, which in turn depends primarily on $p_{\min}$.

Exact analysis of the behavior of $h(\gamma_{\min},m,p_{\min})$ as a function of $\gamma_{\min}$ does not yield interpretable results, however we gain some insight by considering the case where $\gamma_{\min}$ scales with m: $\gamma_{\min} := \tilde{\gamma}(m-1)$. Then we can consider (S40) as a function of $\tilde{\gamma}$ and $p_{\min}$ alone:

$$h(\tilde{\gamma},p_{\min}) := \log\left(\frac{\tilde{\gamma}+1}{\tilde{\gamma}}\right)(\tilde{\gamma}+p_{\min}). \qquad \text{(S41)}$$

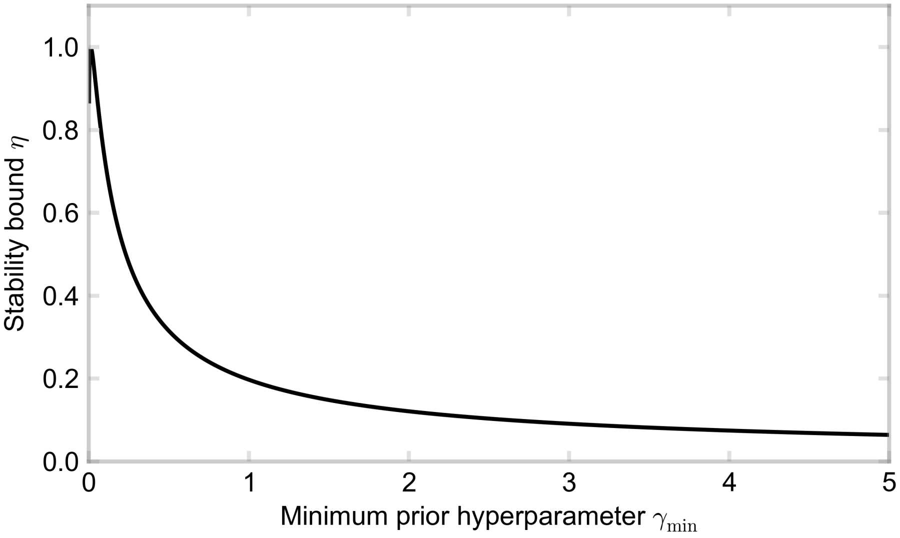

Figure S7: The stability bound η as a function of the prior $\gamma_{\min}$, for fixed $m = 100$ and $p_{\min} = 0.001$. For $\gamma_{\min}$ large enough relative to $p_{\min}$, stronger priors yield tighter bounds.

The bound becomes tighter as $\tilde{\gamma}$ increases, as long as we have $\frac{\partial h(\tilde{\gamma}, p_{\min})}{\partial \tilde{\gamma}} > 0$. This is the case when

$$p_{\min} < \tilde{\gamma}(\tilde{\gamma}+1)\log\left(\frac{\tilde{\gamma}+1}{\tilde{\gamma}}\right) - \tilde{\gamma}. \tag{S42}$$

The quantity on the right-hand side is increasing with $\tilde{\gamma}$. Thus, for $p_{\min}$ small enough relative to $\tilde{\gamma}$, stronger priors lead to a tighter bound. To illustrate this behavior, in Figure S1 we plot the stability bound η (excluding $O\left(\frac{1}{m^2 \log m}\right)$ terms) as a function of $\gamma_{\min}$, for $m = 100$ and $p_{\min} = 0.001$. For $\gamma_{\min}$ larger than about 0.01, the bound tightens as the prior is increased.

3.4 Bayesian Sets and Uniform Stability.

In addition to pointwise hypothesis stability, Bousquet and Elisseeff (2002) define a stronger notion of stability called "uniform stability."

Definition S2 (Bousquet and Elisseeff, 2002)**.** *An algorithm has uniform stability κ with respect to the loss function ℓ if the following holds*

$$\forall S, \quad \forall i \in \{1, \ldots, m\}, \quad ||\ell(f_S, \cdot) - \ell(f_{S^{\setminus i}}, \cdot)||_\infty \leq \kappa. \tag{S43}$$

The algorithm is said to be stable if κ scales with $\frac{1}{m}$.

Uniform stability requires a $O\left(\frac{1}{m}\right)$ bound for all training sets, rather than the average training set as with pointwise hypothesis stability. The bound must also hold for all possible test points, rather than testing on the perturbed point. Uniform stability is actually a very strong condition that is difficult to meet, since if (S43) can be violated by any possible combination of training set and test point, then uniform stability does not hold. Bayesian Sets does not have this form of stability, as we now show with an example.

Choose the training set of m data points to satisfy:

$$\begin{aligned} x_j^i &= 0 \quad \forall j, \quad i = 1, \ldots, m-1 \\ x_j^m &= 1 \quad \forall j, \end{aligned}$$

and as a test point x, take $x_j = 1 \; \forall j$. Let x^m be the point removed from the training set. Then,

$$\begin{aligned}
\kappa &= |\ell(f_S, x) - \ell(f_{S^{\setminus m}}, x)| \\
&= |f_{S^{\setminus m}}(x) - f_S(x)| \\
&= \left| \frac{1}{Z(m-1)} \sum_{j=1}^{N} x_j \log \frac{\alpha_j + \sum_{s=1}^{m} x_j^s - x_j^m}{\alpha_j} - \frac{1}{Z(m)} \sum_{j=1}^{N} x_j \log \frac{\alpha_j + \sum_{s=1}^{m} x_j^s}{\alpha_j} \right| \\
&= \left| \frac{1}{Z(m-1)} \sum_{j=1}^{N} \log \frac{\alpha_j}{\alpha_j} - \frac{1}{Z(m)} \sum_{j=1}^{N} \log \frac{\alpha_j + 1}{\alpha_j} \right| \\
&= \frac{1}{Z(m)} \sum_{j=1}^{N} \log \frac{\alpha_j + 1}{\alpha_j} \\
&\geq \frac{\log \frac{\max_j \alpha_j + 1}{\max_j \alpha_j}}{\log \left(\frac{\gamma_{\min} + m}{\gamma_{\min}} \right)}, && \text{(S44)}
\end{aligned}$$

which scales with m as $\frac{1}{\log m}$, not the $\frac{1}{m}$ required for stability.

References

Bousquet, Olivier and Elisseeff, Andre. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Heller, Katherine A. and Ghahramani, Zoubin. A simple bayesian framework for content-based image retrieval. In *Proceedings of CVPR*, 2006.

Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Johnson, Norman Lloyd, Kemp, Adrienne W., and Kotz, Samuel. *Univariate Discrete Distributions*. John Wiley & Sons, August 2005.

Romanovsky, V. Note on the moments of a binomial $(p+q)^n$ about its mean. *Biometrika*, 15:410–412, 1923.