Supplement to Growing a List

Benjamin Letham* Cynthia Rudin† Katherine A. Heller‡

This supplementary material expands on the experiments and theory given in the main text of Growing a List. In Section 1 we give further detail on the Wikipedia gold standard experiments. In Section 2 we give the proofs of our main theoretical results, Proposition 1 and Theorem 1.

1 Wikipedia Gold Standard Experiments

In Table S1 we give a complete enumeration of the results from the Wikipedia gold standard experiments. For each list growing problem, we provide the Precision@10 and average precision (AveP) 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. We focused on Precision@10 because 10 is the typical number of search results returned by a search engine. We supplement these results further with Precision@5 and Precision@20 in Figure S1.

2 Proofs

In this section, we provide the proofs of Proposition 1 and Theorem 1, comments on the effect of the prior ($\gamma_{\min}$) on generalization, and an example showing that Bayesian Sets does not satisfy the requirements for "uniform stability" defined by Bousquet and Elisseeff (2002).

Recall the definition of the scoring function:

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

where

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

*Operations Research Center, MIT
†MIT Sloan School of Management
‡Center for Cognitive Neuroscience, Statistical Science, Duke

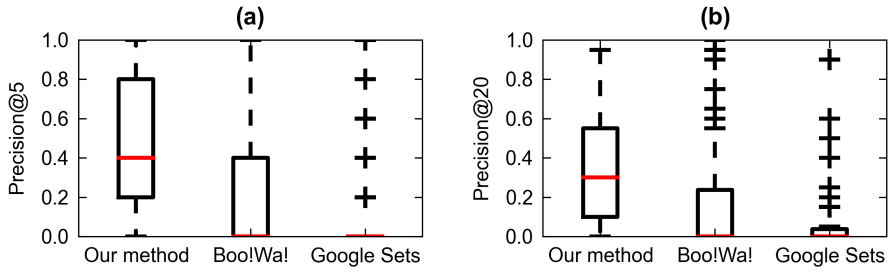

Figure S1: **(a)** Precision@5 and **(b)** Precision@20 across all 50 list growing problems sampled from Wikipedia. The median is indicated in red.

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@10			AveP		
Wikipedia gold standard list	**Us**	**BW**	**GS**	**Us**	**BW**	**GS**
Awards and nominations received by Chris Brown	1	1	0	0.66	0.34	0
Medal of Honor recipients educated at the United States Military Academy	0.2	0	0	0.28	0.01	0
Nine Inch Nails concert tours	0.8	0	0	0.51	0	0
Bleach episodes (season 4)	0	0	0	0	0	0
Storms in the 2005 Atlantic hurricane season	0.1	0	0	0.13	0.11	0
Houses and associated buildings by John Douglas	0.6	0.6	0	0.26	0.32	0
Kansas Jayhawks head football coaches	0.9	0.8	0	0.91	0.79	0
Kraft Nabisco Championship champions	0	0	0	0.05	0.05	0
Washington state symbols	0	0	0	0	0	0
World Heritage Sites of the United Kingdom	0.3	0	0	0.19	0.08	0
Philadelphia Eagles head coaches	0	0	0	0.09	0	0
Los Angeles Dodgers first-round draft picks	0.6	0	0	0.19	0.28	0.00
New York Rangers head coaches	0.3	0.8	0	0.16	0.46	0
African-American Medal of Honor recipients	1	0	0	0.73	0.06	0
Current sovereign monarchs	0.5	0	0	0.15	0	0
Brotherhood episodes	0.9	0.2	0	0.72	0.06	0
Knight's Cross of the Iron Cross with Oak Leaves recipients (1945)	0	0	0	0	0.01	0.00
Pittsburgh Steelers first-round draft picks	0.1	0	0	0.38	0.00	0
Tallest buildings in New Orleans	0.6	0	0.6	0.45	0	0.08
Asian XI ODI cricketers	0.2	0	0.4	0.18	0.01	0.08
East Carolina Pirates head football coaches	0.1	0	0	0.05	0.01	0
Former championships in WWE	0.4	0	0.4	0.31	0.09	0.15
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	0.4	0	0.2	0.27	0.14	0.02
Wilfrid Laurier University people	0.3	0	0	0.34	0.11	0
Wario video games	0.1	0.6	0.8	0.12	0.22	0.34
Governors of Washington	0.5	0	0	0.42	0.13	0
Buffalo Sabres players	0.1	0	0	0.03	0	0
Australia Twenty20 International cricketers	0.6	0	1	0.24	0.01	0.32
Awards and nominations received by Madonna	0.9	1	0.2	0.70	0.13	0.00
Yukon Quest competitors	0.7	0.4	0.2	0.02	0.35	0.00
Arsenal F.C. players	0.8	0	0	0.85	0.18	0
Victoria Cross recipients of the Royal Navy	0.4	0	0	0.12	0.01	0
Formula One drivers	0	0.6	1	0	0.15	0.01
Washington & Jefferson College buildings	0	0	0	0	0	0
X-Men video games	0.4	0.2	0	0.27	0.05	0
Governors of Florida	0.4	0	0	0.25	0.04	0
The Simpsons video games	0.1	0	0	0.18	0	0
Governors of New Jersey	0.7	0	0	0.34	0.07	0
Uncharted characters	0.4	0	0.6	0.27	0.01	0.33
Miami Marlins first-round draft picks	0.4	1	0	0.16	0.27	0
Tallest buildings in Dallas	0.7	0.2	0	0.34	0.14	0
Cities and towns in California	0.8	0.6	1	0.35	0.04	0.04
Olympic medalists in badminton	0.3	0	0	0.13	0.05	0
Delegates to the Millennium Summit	0.9	0.4	0	0.51	0.01	0
Honorary Fellows of Jesus College, Oxford	0	0.4	0	0.03	0.34	0
Highlander: The Raven episodes	0.2	1	0	0.14	0.95	0
Voice actors in the Grand Theft Auto series	0.4	0	0	0.16	0.18	0
Medal of Honor recipients for the Vietnam War	0.9	0.8	0	0.84	0.08	0

and $\gamma_{\min} := \min_j \min\{\alpha_j, \beta_j\}$. We begin by showing that the normalized score $f_S(x)$ in (S1) takes values only on $[0, 1]$.

Lemma S1. $0 \le f_S(x) \le 1$.

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

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

□

Now we provide the proof to Proposition 1.

Proof of Proposition 1. 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,

$$\begin{aligned}
&\frac{1}{m} \sum_{s=1}^{m} f_S(x^s) - \mathbb{E}_x \left[f_S(x) \right] \\
&\qquad = \frac{1}{N \log \left(\frac{\gamma_{\min} + m}{\gamma_{\min}} \right)} \sum_{j=1}^{N} \left((\mu_j - p_j) \log \frac{\alpha_j + m\mu_j}{\alpha_j} + (p_j - \mu_j) \log \frac{\beta_j + m(1 - \mu_j)}{\beta_j} \right) \\
&\qquad \le \frac{1}{N} \sum_{j=1}^{N} |\mu_j - p_j|. \qquad \text{(S2)}
\end{aligned}$$

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

$$\mathbb{P}(|\mu_j - p_j| > \epsilon) \le 2 \exp\left(-2m\epsilon^2\right). \qquad \text{(S3)}$$

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

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{N} \sum_{j=1}^{N} |\mu_j - p_j| > \epsilon \right) &\le \mathbb{P}\left(\bigcup_{j=1}^{N} \{|\mu_j - p_j| > \epsilon\} \right) \\
&\le \sum_{j=1}^{N} \mathbb{P}\left(|\mu_j - p_j| > \epsilon \right) \\
&\le 2N \exp\left(-2m\epsilon^2\right). \qquad \text{(S4)}
\end{aligned}$$

Thus,

$$\mathbb{P}\left(\frac{1}{m}\sum_{s=1}^{m} f_S(x^s) - \mathbb{E}_x\left[f_S(x)\right] > \epsilon\right) \leq 2N\exp\left(-2m\epsilon^2\right), \tag{S5}$$

and the proposition follows directly. □

The bound in Proposition 1 has a tighter dependence on δ than the bound in Theorem 1, however it depends inversely on N, the number of features.

We now present the proof of Theorem 1. The result uses the algorithmic stability bounds of Bousquet and 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:

$$\ell(f_S, x) = 1 - f_S(x). \tag{S6}$$

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: $S^{\setminus i} := \{x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^m\}$. We then define the score of x according to the modified training set:

$$f_{S^{\setminus i}}(x) = \frac{1}{Z(m-1)}\sum_{j=1}^{N} x_j \log\frac{\alpha_j + \sum_{s\neq i} x_j^s}{\alpha_j} + (1-x_j)\log\frac{\beta_j + (m-1) - \sum_{s\neq i} x_j^s}{\beta_j}, \tag{S7}$$

where

$$Z(m-1) = N\log\left(\frac{\gamma_{\min} + m - 1}{\gamma_{\min}}\right). \tag{S8}$$

We further define the loss using the modified training set:

$$\ell(f_{S^{\setminus i}}, x) = 1 - f_{S^{\setminus i}}(x). \tag{S9}$$

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 and 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 and Elisseeff, 2002)**.** *An algorithm has pointwise hypothesis stability η with respect to the loss function ℓ if the following holds*

$$\forall i \in \{1, \ldots, m\}, \quad \mathbb{E}_S\left[\left|\ell(f_S, x^i) - \ell(f_{S^{\setminus i}}, x^i)\right|\right] \leq \eta. \tag{S10}$$

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 and 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 and Elisseeff (2002), with the notation adapted to our setting.

Theorem S1 (Bousquet and 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 $1-\delta$,*

$$\mathbb{E}_x\left[\ell(f_S, x)\right] \leq \frac{1}{m}\sum_{i=1}^{m}\ell(f_S, x^i) + \sqrt{\frac{1 + 12m\eta}{2m\delta}}. \tag{S11}$$

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 S1. 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):

$$\mu_{s+1} = p(1-p)\left(\frac{d\mu_s}{dp} + ms\mu_{s-1}\right). \tag{S12}$$

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 (S12),

$$\mu_{2(s+1)} = p(1-p)\left(\frac{d\mu_{2s+1}}{dp} + (2s+1)m\mu_{2s}\right). \tag{S13}$$

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,

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

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 $\alpha > 0$,*

$$\mathbb{E}\left[\frac{1}{\alpha + t}\right] = \frac{1}{\alpha + mp} + O\left(\frac{1}{m^2}\right). \tag{S15}$$

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

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\alpha + t}\right] &= \mathbb{E}\left[\sum_{i=0}^{\infty}(-1)^i \frac{(t-mp)^i}{(\alpha+mp)^{i+1}}\right] \\
&= \sum_{i=0}^{\infty}(-1)^i \frac{\mathbb{E}\left[(t-mp)^i\right]}{(\alpha+mp)^{i+1}} \\
&= \frac{1}{\alpha+mp} + \sum_{i=2}^{\infty}(-1)^i \frac{\mu_i}{(\alpha+mp)^{i+1}}
\end{aligned} \tag{S16}$$

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

$$\frac{\mu_i}{(\alpha+mp)^{i+1}} = \frac{O\left(m^{\lfloor \frac{i}{2} \rfloor}\right)}{O\left(m^{i+1}\right)} = O\left(m^{\lfloor \frac{i}{2} \rfloor - i - 1}\right). \tag{S17}$$

The alternating sum in (S16) can be split into two sums:

$$\sum_{i=2}^{\infty}(-1)^i \frac{\mu_i}{(\alpha+mp)^{i+1}} = \sum_{i=2}^{\infty} O\left(m^{\lfloor \frac{i}{2} \rfloor - i - 1}\right) = \sum_{i=2}^{\infty} O\left(\frac{1}{m^i}\right) + \sum_{i=3}^{\infty} O\left(\frac{1}{m^i}\right). \tag{S18}$$

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 $\alpha > 0$,*

$$g(\alpha, m) := \frac{\log\left(\frac{\alpha+m}{\alpha}\right)}{\log\left(\frac{\alpha+m-1}{\alpha}\right)} \tag{S19}$$

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 (S19) to

$$g(a, b) := \frac{\log(ab+1)}{\log(a+1)}. \tag{S20}$$

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

$$h(a, b) := (ab+1)\log(ab+1) - b(a+1)\log(a+1) \geq 0. \tag{S21}$$

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

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

because $b \geq 1$ and $a \geq \log(1+a)\ \forall a \geq 0$. This shows that (S21) 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 $\gamma_{\min} \in (0, \alpha]$,*

$$\frac{1}{Z(m)}\log\frac{\alpha+t+1}{\alpha} \geq \frac{1}{Z(m-1)}\log\frac{\alpha+t}{\alpha}. \tag{S22}$$

Proof. Denote,

$$g(t; m, \alpha) := \frac{1}{Z(m)}\log\frac{\alpha+t+1}{\alpha} - \frac{1}{Z(m-1)}\log\frac{\alpha+t}{\alpha}. \tag{S23}$$

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,

$$\frac{\log\left(\frac{\alpha+m}{\alpha}\right)}{Z(m)} \geq \frac{\log\left(\frac{\alpha+m-1}{\alpha}\right)}{Z(m-1)}, \tag{S24}$$

which shows

$$g(m-1; m, \alpha) = \frac{1}{Z(m)}\log\frac{\alpha+m}{\alpha} - \frac{1}{Z(m-1)}\log\frac{\alpha+m-1}{\alpha} \geq 0. \tag{S25}$$

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

$$\frac{\partial g(t; m, \alpha)}{\partial t} = \frac{1}{Z(m)}\frac{1}{\alpha+t+1} - \frac{1}{Z(m-1)}\frac{1}{\alpha+t} < 0, \tag{S26}$$

for all $t \geq 0$. Equations S25 and S26 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 $\gamma_{\min} \in (0, \beta]$,*

$$\frac{1}{Z(m)} \log \frac{\beta + m - t}{\beta} \geq \frac{1}{Z(m-1)} \log \frac{\beta + m - 1 - t}{\beta}. \tag{S27}$$

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

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

□

The next lemma is the key lemma that shows Bayesian Sets satisfies pointwise hypothesis stability, allowing us to apply Theorem S1.

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

$$\eta = \frac{1}{\log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)(\gamma_{\min} + (m-1)p_{\min})} + O\left(\frac{1}{m^2 \log m}\right). \tag{S29}$$

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| \qquad \text{(S30)} \\
&\quad := \mathbb{E}_S \sum_{j=1}^{N} x_j^i \mathrm{term}_j^1 + (1 - x_j^i) \mathrm{term}_j^2 \qquad \text{(S31)} \\
&\quad = \sum_{j=1}^{N} \mathbb{E}_{x_j^1, \ldots, x_j^m} \left[x_j^i \mathrm{term}_j^1 + (1 - x_j^i) \mathrm{term}_j^2 \right] \\
&\quad = \sum_{j=1}^{N} \mathbb{E}_{x_j^i} \left[\mathbb{E}_{x_j^{s \neq i} | x_j^i} \left[x_j^i \mathrm{term}_j^1 \right] \right] + \mathbb{E}_{x_j^i} \left[\mathbb{E}_{x_j^{s \neq i} | x_j^i} \left[(1 - x_j^i) \mathrm{term}_j^2 \right] \right] \\
&\quad = \sum_{j=1}^{N} \mathbb{E}_{x_j^i} \left[x_j^i \mathbb{E}_{x_j^{s \neq i} | x_j^i} \left[\mathrm{term}_j^1 \right] \right] + \mathbb{E}_{x_j^i} \left[(1 - x_j^i) \mathbb{E}_{x_j^{s \neq i} | x_j^i} \left[\mathrm{term}_j^2 \right] \right] \\
&\quad = \sum_{j=1}^{N} \mathbb{E}_{x_j^{s \neq i}} \left[\mathrm{term}_j^1 | x_j^i = 1 \right] \mathbb{P}\left(x_j^i = 1 \right) + \mathbb{E}_{x_j^{s \neq i}} \left[\mathrm{term}_j^2 | x_j^i = 0 \right] \mathbb{P}\left(x_j^i = 0 \right) \\
&\quad \leq \sum_{j=1}^{N} \max \left\{ \mathbb{E}_{x_j^{s \neq i}} \left[\mathrm{term}_j^1 | x_j^i = 1 \right], \mathbb{E}_{x_j^{s \neq i}} \left[\mathrm{term}_j^2 | x_j^i = 0 \right] \right\}, \qquad \text{(S32)}
\end{aligned}$$

where (S30) uses the triangle inequality, and in (S31) we define term_j^1 and term_j^2 for notational convenience. Now consider each term in (S32) separately,

$$\begin{aligned}\mathbb{E}_{x_j^{s\neq i}}\left[\text{term}_j^1 | x_j^i = 1\right] &= \mathbb{E}_{x_j^{s\neq i}}\left|\frac{1}{Z(m-1)}\log\frac{\alpha_j + \sum_{s\neq i} x_j^s}{\alpha_j} - \frac{1}{Z(m)}\log\frac{\alpha_j + \sum_{s\neq i} x_j^s + 1}{\alpha_j}\right| \\ &= \mathbb{E}_{x_j^{s\neq i}}\left[\frac{1}{Z(m)}\log\frac{\alpha_j + \sum_{s\neq i} x_j^s + 1}{\alpha_j} - \frac{1}{Z(m-1)}\log\frac{\alpha_j + \sum_{s\neq i} x_j^s}{\alpha_j}\right], \end{aligned} \tag{S33}$$

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 (S33),

$$\begin{aligned}\mathbb{E}_{x_j^{s\neq i}}\left[\text{term}_j^1 | x_j^i = 1\right] &= \mathbb{E}_{t\sim\text{Bin}(m-1,p_j)}\left[\frac{1}{Z(m)}\log\frac{\alpha_j + t + 1}{\alpha_j} - \frac{1}{Z(m-1)}\log\frac{\alpha_j + t}{\alpha_j}\right] \\ &\leq \frac{1}{Z(m-1)}\mathbb{E}_{t\sim\text{Bin}(m-1,p_j)}\left[\log\frac{\alpha_j + t + 1}{\alpha_j + t}\right] \\ &= \frac{1}{Z(m-1)}\mathbb{E}_{t\sim\text{Bin}(m-1,p_j)}\left[\log\left(1 + \frac{1}{\alpha_j + t}\right)\right] \\ &\leq \frac{1}{Z(m-1)}\log\left(1 + \mathbb{E}_{t\sim\text{Bin}(m-1,p_j)}\left[\frac{1}{\alpha_j + t}\right]\right) \\ &= \frac{1}{Z(m-1)}\log\left(1 + \frac{1}{\alpha_j + (m-1)p_j} + O\left(\frac{1}{m^2}\right)\right). \end{aligned} \tag{S34}$$

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.

$$\begin{aligned}&\mathbb{E}_{x_j^{s\neq i}}\left[\text{term}_j^2 | x_j^i = 0\right] \\ &\quad= \mathbb{E}_{x_j^{s\neq i}}\left|\frac{1}{Z(m-1)}\log\frac{\beta_j + (m-1) - \sum_{s\neq i} x_j^s}{\beta_j} - \frac{1}{Z(m)}\log\frac{\beta_j + m - \sum_{s\neq i} x_j^s}{\beta_j}\right| \\ &\quad= \mathbb{E}_{x_j^{s\neq i}}\left[\frac{1}{Z(m)}\log\frac{\beta_j + m - \sum_{s\neq i} x_j^s}{\beta_j} - \frac{1}{Z(m-1)}\log\frac{\beta_j + (m-1) - \sum_{s\neq i} x_j^s}{\beta_j}\right]. \end{aligned} \tag{S35}$$

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 (S35):

$$\begin{aligned}\mathbb{E}_{x_j^{s\neq i}}\left[\text{term}_j^2 | x_j^i = 0\right] &\leq \frac{1}{Z(m-1)}\mathbb{E}_{t\sim\text{Bin}(m-1,q_j)}\left[\log\frac{\beta_j + t + 1}{\beta_j + t}\right] \\ &\leq \frac{1}{Z(m-1)}\log\left(1 + \frac{1}{\beta_j + (m-1)q_j} + O\left(\frac{1}{m^2}\right)\right). \end{aligned} \tag{S36}$$

where the steps are as with (S34). We now take (S34) and (S36) and use them to continue (S32):

$$\begin{aligned}
&\mathbb{E}_S|\ell(f_S, x^i) - \ell(f_{S^{\setminus i}}, x^i)| \\
&\quad\leq \sum_{j=1}^{N} \max\left\{ \frac{1}{Z(m-1)} \log\left(1 + \frac{1}{\alpha_j + (m-1)p_j} + O\left(\frac{1}{m^2}\right)\right), \right. \\
&\qquad\qquad \left. \frac{1}{Z(m-1)} \log\left(1 + \frac{1}{\beta_j + (m-1)q_j} + O\left(\frac{1}{m^2}\right)\right)\right\} \\
&\quad\leq \sum_{j=1}^{N} \frac{1}{Z(m-1)} \log\left(1 + \frac{1}{\min\{\alpha_j, \beta_j\} + (m-1)\min\{p_j, q_j\}} + O\left(\frac{1}{m^2}\right)\right) \\
&\quad\leq \frac{N}{Z(m-1)} \log\left(1 + \frac{1}{\gamma_{\min} + (m-1)p_{\min}} + O\left(\frac{1}{m^2}\right)\right) \\
&\quad:= \eta.
\end{aligned} \tag{S37}$$

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

$$\begin{aligned}
\eta &= \frac{N}{Z(m-1)} \left(\frac{1}{\gamma_{\min} + (m-1)p_{\min}} + O\left(\frac{1}{m^2}\right) - \frac{1}{2}\left(\frac{1}{\gamma_{\min} + (m-1)p_{\min}} + O\left(\frac{1}{m^2}\right)\right)^2 \right) \\
&= \frac{N}{Z(m-1)} \left(\frac{1}{\gamma_{\min} + (m-1)p_{\min}} + O\left(\frac{1}{m^2}\right)\right) \\
&= \frac{1}{\log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)(\gamma_{\min} + (m-1)p_{\min})} + O\left(\frac{1}{m^2 \log m}\right).
\end{aligned} \tag{S38}$$

□

The proof of Theorem 1 is now a straightforward application of Theorem S1 using the result of Lemma S7.

Proof of Theorem 1. By Lemma S7, we can apply Theorem S1 to see that with probability at least $1-\delta$ on the draw of S,

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

□

2.1 Comments on the effect of the prior on generalization.

The prior influences the generalization bound via the quantity

$$h(\gamma_{\min}, m, p_{\min}) := \log\left(\frac{\gamma_{\min}+m-1}{\gamma_{\min}}\right)(\gamma_{\min} + (m-1)p_{\min}). \tag{S39}$$

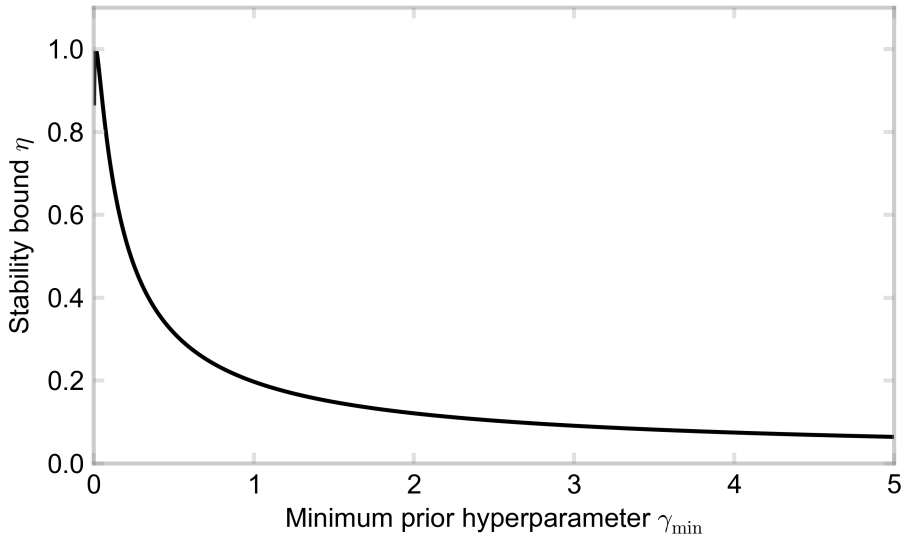

Figure S2: The stability bound η as a function of the prior $\gamma_{\min}$, for fixed $m = 100$ and $p_{\min} = 0.001$. For $\gamma_{\min}$ large enough relative to $p_{\min}$, stronger priors yield tighter bounds.

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 (S39) as a function of $\tilde{\gamma}$ and $p_{\min}$ alone:

$$h(\tilde{\gamma}, p_{\min}) := \log\left(\frac{\tilde{\gamma}+1}{\tilde{\gamma}}\right)(\tilde{\gamma} + p_{\min})\,. \tag{S40}$$

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

$$p_{\min} < \tilde{\gamma}(\tilde{\gamma}+1)\log\left(\frac{\tilde{\gamma}+1}{\tilde{\gamma}}\right) - \tilde{\gamma}. \tag{S41}$$

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.

2.2 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*

$$\forall S, \quad \forall i \in \{1, \ldots, m\}, \quad ||\ell(f_S, \cdot) - \ell(f_{S^{\setminus i}}, \cdot)||_\infty \leq \kappa. \tag{S42}$$

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 (S42) 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,

$$\begin{aligned} \kappa &= |\ell(f_S, x) - \ell(f_{S^{\setminus m}}, x)| \\ &= |f_{S^{\setminus m}}(x) - f_S(x)| \\ &= \left| \frac{1}{Z(m-1)} \sum_{j=1}^N x_j \log \frac{\alpha_j + \sum_{s=1}^m x_j^s - x_j^m}{\alpha_j} - \frac{1}{Z(m)} \sum_{j=1}^N x_j \log \frac{\alpha_j + \sum_{s=1}^m x_j^s}{\alpha_j} \right| \\ &= \left| \frac{1}{Z(m-1)} \sum_{j=1}^N \log \frac{\alpha_j}{\alpha_j} - \frac{1}{Z(m)} \sum_{j=1}^N \log \frac{\alpha_j + 1}{\alpha_j} \right| \\ &= \frac{1}{Z(m)} \sum_{j=1}^N \log \frac{\alpha_j + 1}{\alpha_j} \\ &\geq \frac{\log \frac{\max_j \alpha_j + 1}{\max_j \alpha_j}}{\log\left(\frac{\gamma_{\min} + m}{\gamma_{\min}}\right)}, \end{aligned} \tag{S43}$$

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

References

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

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

Johnson NL, Kemp AW, Kotz S (2005) Univariate Discrete Distributions. John Wiley & Sons

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