Estimation of Large Zipfian Distributions with Sort and Snap

Peter Jacobs [1, 2] Debdeep Pati [3] Anirban Bhattacharya [4] Lekha Patel [2] Jeff Phillips [1]

[1]University of Utah [3]University of Wisconsin [4]Texas A&M University [2]Sandia National Laboratories

What is a Zipfian Distribution?

- A discrete distribution ($\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$) over k alphabet items where for some decay $s > 0$ and permutation function π, the i^{th} largest probability, denoted $p_{\pi(i)}$, satisfies

$$p_{\pi(i)} \propto i^{-s}$$

- $s > 0$ is called the decay. π gives the order of the alphabet items in decreasing order of probability

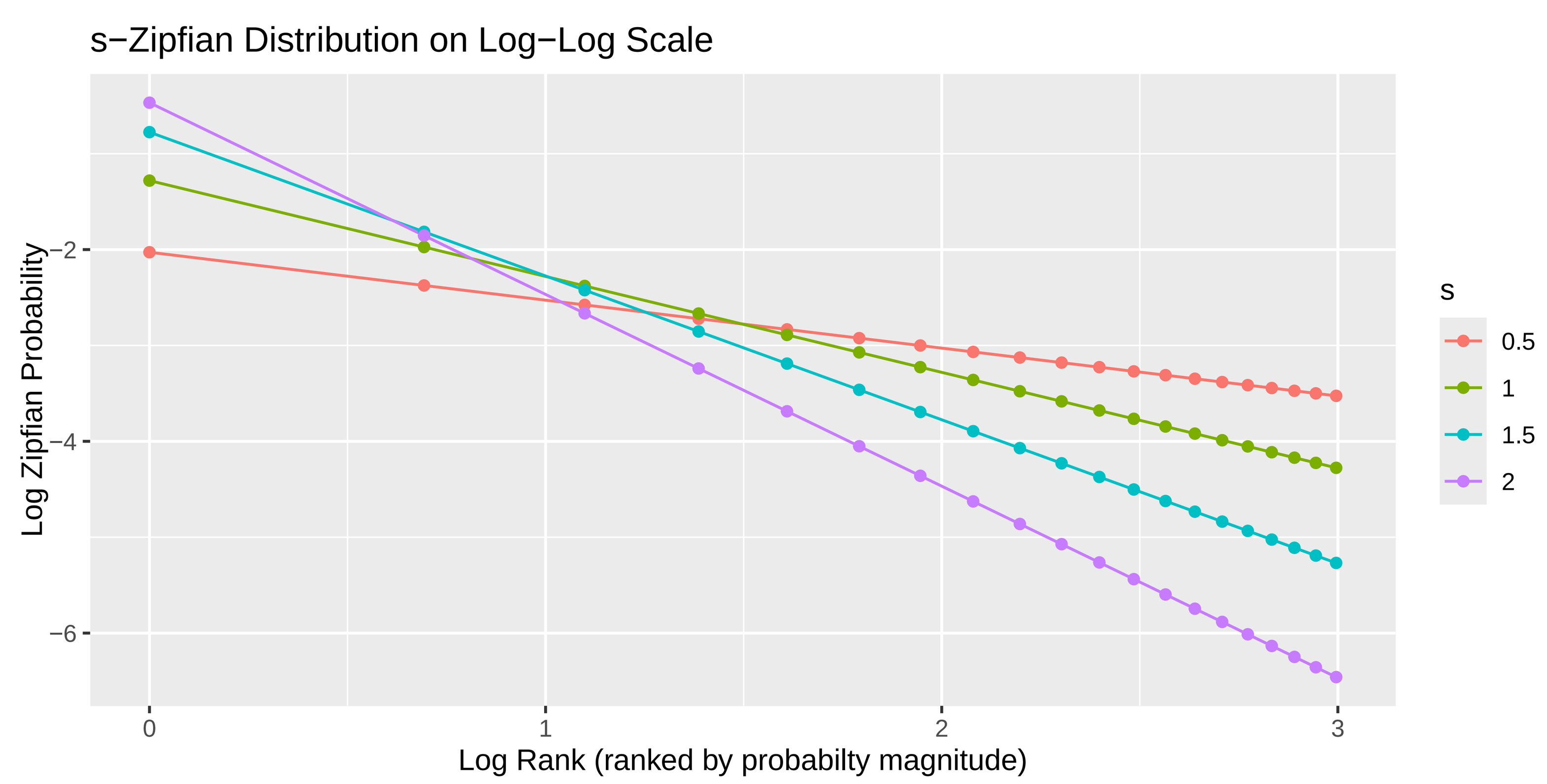

Figure 1. The top 20 probabilities in decreasing order for the $k = 100$ and $s \in \{.5, 1, 1.5, 2\}$ Zipfian distributions

What Problem Do We Consider?

1. $\boldsymbol{p}$ is s-Zipfian for some $s > 0, k \in \mathbb{N}$ and some permutation function π.
2. We must learn $\boldsymbol{p}$ only from n i.i.d samples $Y_1, Y_2, \ldots, Y_n$
3. High Dimensional Inference: $k = n^\beta$ for some $\beta > 0$.
4. We seek an *estimator* for $\boldsymbol{p}$, denoted $\hat{\boldsymbol{p}}$, that is close to $\boldsymbol{p}$ for large n

How Do We Compare Estimators?

Let $\mathcal{P}_{s,k}$ denote the k-dimensional $s > 0$ Zipfian distributions

- A *good* estimator $\hat{p}$ will minimize maximum risk over $\mathcal{P}_{s,k}$ as n grows:

$$\sup_{\boldsymbol{p} \in \mathcal{P}_{s,k}} \mathbb{E}_{\boldsymbol{p}} L_1(\hat{\boldsymbol{p}}, \boldsymbol{p}) \approx \inf_{\hat{\boldsymbol{p}} \in \text{Estimators}} \sup_{\boldsymbol{p} \in \mathcal{P}_{s,k}} \mathbb{E}_{\boldsymbol{p}} L_1(\hat{\boldsymbol{p}}, \boldsymbol{p})$$

- For $\boldsymbol{p}, \boldsymbol{q}$, $L_1(\boldsymbol{p}, \boldsymbol{q}) := \sum_{j=1}^k |p_i - q_i|$

Why is this problem important?

Because high-dimensional Zipfian or near Zipfian distributions occur in a variety of natural settings. Examples include City Size and Word Frequency distributions.

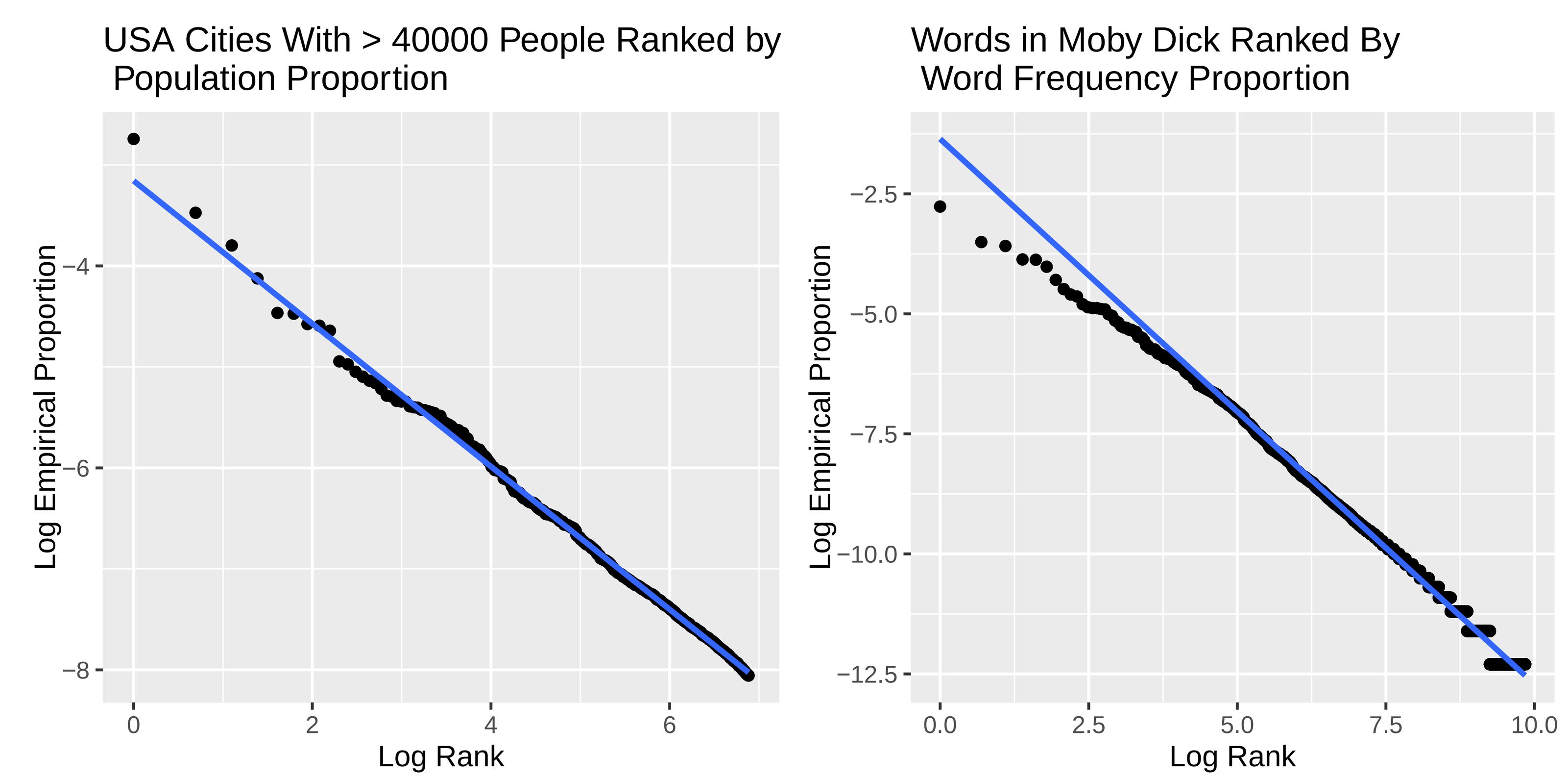

Figure 2. Double Log Plots for City Size Proportions in the USA (left,$n = 125,145,895$ people,$k = 976$ cities) and Word Frequency Proportions in Moby Dick (right,$n = 219,422$,$k = 18,811$ words). The Blue line represents a Zipfian Probability Law fit using Ordinary Least Squares

Our Approach

Is s known before data collection?

- If Yes: Use **Sort and Snap**: Sort the alphabet items by empirical frequency, and assign the i^{th} largest s-Zipfian probability as the estimate for the alphabet item with i^{th} largest frequency
- If No: **Adaptive Sort and Snap**: Estimate s first via approximate Maximum Likelihood. Then apply Sort and Snap with given estimate of s.

Through theory and experiment, we compare **Sort and Snap** (SS) and **Adaptive Sort and Snap** (SSA) to SOTA (state of the art) estimators.

Estimator	Usable for s unknown?
Sort and Snap (SS)	No
Adaptive Sort and Snap (SSA)	Yes
Empirical Proportions (EPE)	Yes
Good-Turing (EPE-GT) [3]	Yes
Absolute Discounting (AD) [2]	Yes
Braess-Sauer (BS) [1]	Yes

Table 1. Estimators Under Comparison

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525

Results

Highlighted Theoretical Results

In the s known case:

- SS is near minimax and outperforms SOTA for $\beta < \frac{1}{B(s)}$ where $B(s) := \max(1, s) + 2$
- SS achieves an *exponential decay* in worst case risk when $\beta < \frac{1}{B(s)}$ of general form $n^{-f(\beta,s)} \exp(-n^{1-\beta B(s)})$
- An adaptation of SS that uses SS only for top $\frac{1}{B(s)}$ ranks and EPE otherwise is near minimax when $\beta > \frac{1}{B(s)}$ and $s > 2$

Here we highlight experiments in which the worst case risk of the estimators is compared at various (s, β).

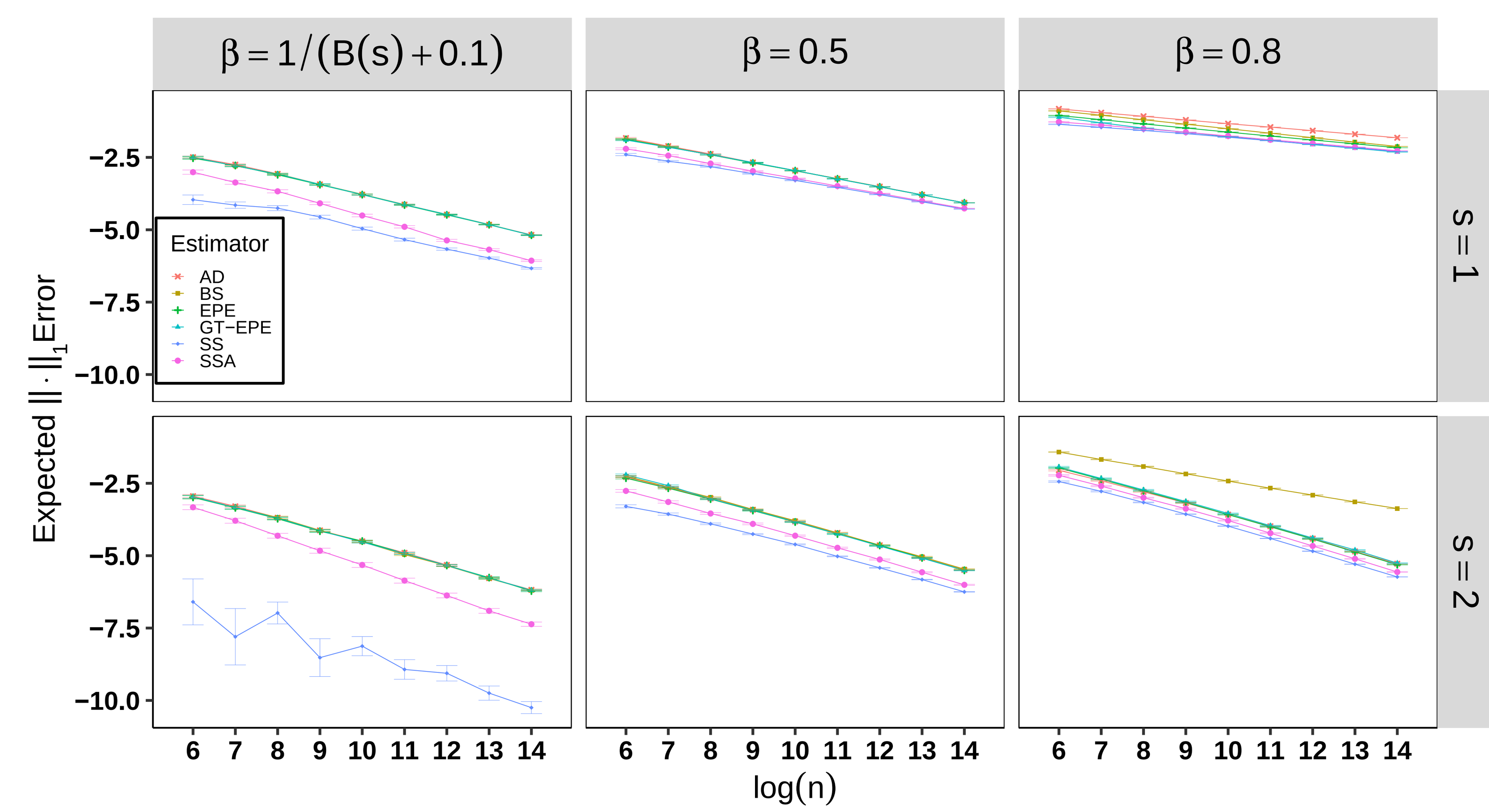

Figure 3. Monte Carlo Study Using $M = 300$ iterations to compare expected risk under a variety of (s, β) settings. Note $B(s) := \max(s, 1) + 2$, which is an intermediate dimension at which Sort and Snap no longer has exponential decay in its risk

Summary of Conclusions

- The above simulations, and our theoretical results suggest SS and SSA are competitive with SOTA methods for all (s, β), and superior when β is small relative to s.
- Future work will investigate
  - SSA for more complex distriutions (Zipf-Mandelbrot) that occur in nature
  - How to incorporate SSA into a model-selection strategy when the model is unknown.