Imposing Monotonicity Improves Large-Alphabet Distribution Estimation

Amit Geva Amichai Painsky
Department of Statistics and Operations Research, Tel Aviv University

Motivation

Large-Alphabet Distribution Estimation

Estimating probability distributions over large alphabets is a fundamental problem in statistics, machine learning, and information theory. In many modern applications, the alphabet size is comparable to or larger than the sample size, resulting in many rare and unobserved events.

Classical estimators such as:

- Good-Turing (GT)
- Gale and Sampson (GS)
- Orlitsky and Suresh (OS)

were specifically designed to address this setting.

However, these estimators often violate a natural structural property: symbols observed more frequently should receive higher estimated probabilities.

Example: Non-Monotonicity of the Good–Turing Estimator

Consider the sample of 8 observations

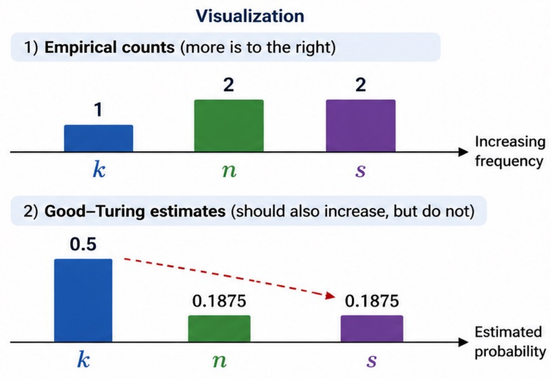
$$X^n = \{k, e, e, n, n, e, s, s\}$$

- k appears 1 time
- n and s appear 2 times each

Empirical counts and Good–Turing estimates

Symbol	Empirical Count $N_u(X^n)$	Good–Turing Estimate $\hat{p}_{GT}(u)$
k	1	$\frac{1}{2} = 0.5$
n	2	$\frac{3}{16} = 0.1875$
s	2	$\frac{3}{16} = 0.1875$

Visualization



This violates monotonicity and may lead to suboptimal estimation performance.

Main Idea

Enforcing Monotonicity via Information Projection

Instead of designing a new estimator, we improve existing estimators using a simple post-processing step based on **information projection (I-projection)**.

We define the set of distributions consistent with the empirical ordering:

$$C(X^n) = \{q \in \Delta_m : q(u) \le q(v)\ \forall\ N_u(X^n) \le N_v(X^n)\}$$

where Δ_m is the distibution simplex, N_u be the number of apperences of the symbols u in X^n.

We then project any estimator $\hat{p}$ onto this set using KL divergence:

$$\hat{p}^* = \arg\min_{q \in C(X^n)} D_{KL}(q||\hat{p})$$

- The resulting estimator $\hat{p}^*$ is the closest monotone distribution to the original estimator under KL divergence.
- The I-projection enforces monotonicity while minimally modifying the original estimate.

Interpretation

The proposed framework:

- Smooths irregular probability estimates
- Acts as implicit regularization
- Improves estimation accuracy

Implementation

The I-projection can be efficiently approximated using **isotonic regression:**

$$\hat{p}^* = \arg\min_{q_1 \le q_2 \le \cdots \le q_m} \sum \hat{p}_i (q_i - \log \hat{p}_i)^2$$

The projected estimates are exponentiated, normalized to sum to one, and remain non-negative so that $\hat{p}^* \in \Delta_m$.

Theoretical Guarantees

Theorem 1 – Improvement in KL Divergence

Let $\hat{p}^*$ denote the I-projection of $\hat{p}$ onto the monotone set $C(X^n)$.

Then:

$$D_{KL}(p||\hat{p}^*) \le D_{KL}(p||\hat{p})$$

with high probability.

Theorem 2 – Risk Improvement

Define the smoothed projection estimator:

$$\tilde{p}^* = (1 - ma)\hat{p}^* + a$$

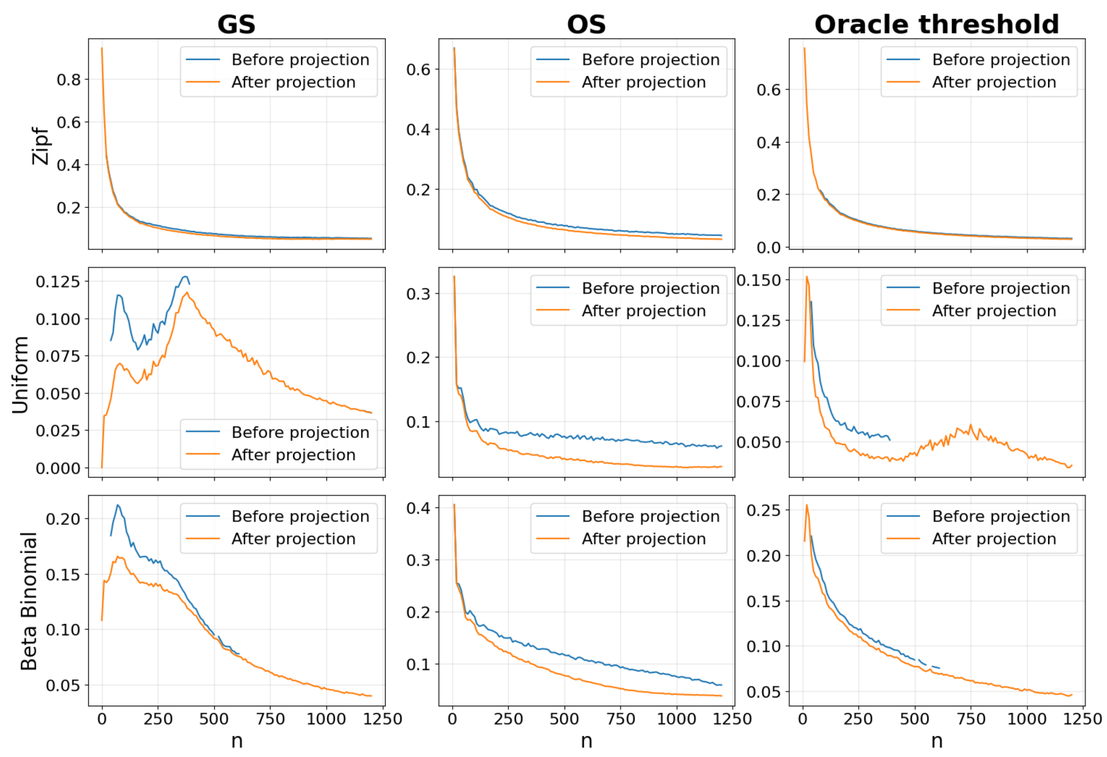
where $a > 0$ is a small smoothing constant ensuring strictly positive probabilities.

Then, under mild conditions on the alphabet growth rate,

$$\mathbb{E}[D_{KL}(p||\tilde{p}^*)] \le \mathbb{E}[D_{KL}(p||\hat{p})] + o(1/n)$$

Experimental Results

- We evaluate the proposed monotone projection framework on four representative synthetic distributions:
  - Zipf's law ($s = 1.01$)
  - Uniform
  - Beta-Binomial ($\alpha = \beta = 2$)
- For each setting, we compare the performance of three large-alphabet estimators before and after applying the monotone projection:
  - Gale and Sampson (GS)
  - Orlitsky and Suresh (OS)
  - An oracle threshold estimator
- Performance is measured using the average Kullback–Leibler divergence over 1000 independent simulations as the sample size n varies.



- The results demonstrate that the proposed projection consistently improves estimation accuracy across nearly all examined settings.
- The improvement is especially pronounced for distributions where the baseline estimators exhibit strong non-monotonic behavior, such as the Uniform and Beta-Binomial cases.
- In contrast, the gains are smaller for Zipf distributions, since many existing estimators are specifically designed for Zipf-like distributions.
- Similar improvements are observed on real-world datasets, including the Hamilton corpus, Biota microbiome dataset, and U.S. Census surnames dataset.