Entropy and Inference, Revisited

Ilya Nemenman,[1,3] Fariel Shafee,[2]
William Bialek[1,2]

[1]NEC Research Institute

[2]Princeton University

[3]University of California, Santa Barbara

We study properties of popular, near–uniform, priors for learning undersampled probability distributions on discrete nonmetric spaces and show that they lead to disastrous results. However, an Occam–style phase space argument allows us to salvage the priors and turn the problems into a surprisingly good estimator of entropies of discrete distributions.

Undersampled learning of probabilities on

continuous spaces (weather, stocks,. . .):

Possible outcomes	$x, a \le x \le b$
Probability density	$Q(x)$
Observed data	$x_\mu,\ \mu = 1 \dots N$
Undersampled regime	always
Smoothness	$\partial^\eta Q / \partial x^\eta$ is small
Regularization of learning	local: punish for $\partial^\eta Q / \partial x^\eta \gg 1$
Model selection	phase space volume, self-consistent
Prior-insensitive learning	probably possible

discrete nonmetric spaces (languages, bioinformatics,. . .):

Discrete outcomes (bins)	$i,\ i = 1 \dots K$
Probability mass	q_i
Observed bin occupancy	n_i
Undersampled regime	$\sum_{i=1}^{K} n_i \equiv N \ll K$
Smoothness	undefined
Regularization of learning	ultralocal: $\mathcal{P}(\{q_i\}) = \prod \mathcal{P}_i(q_i)$ global: $\mathcal{P}(\{q_i\}) = F(\text{entropy})$
Model selection	unknown
Prior-insensitive learning	probably impossible for $N \ll K$

Our options (for discrete case):

1. Define smoothness as high entropy or low mutual information distributions.
2. Prior-insensitive learning of useful functions (like entropy) may be possible for $N \ll K$ even if it's impossible for $\{q_i\}$.

We choose:

Learning entropy with nearly uniform priors

Family of priors: (Dirichlet priors)

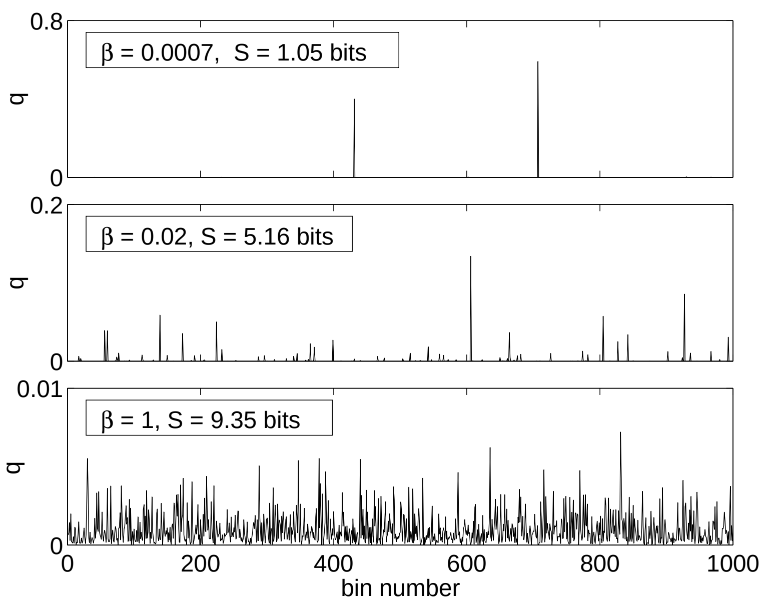
$$\mathcal{P}_\beta(\{q_i\}) = \frac{1}{Z(\beta)}\delta\left(1-\sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^{\beta-1}$$

Generation of distributions from this family:

Successively select each q_i according to

$$\begin{aligned} P(q_i) &= B\left(\frac{q_i}{1-\sum_{j<i} q_j};\beta,(K-i)\beta\right)\\ B\left(x;a,b\right) &= \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}\end{aligned}$$

Typical distributions ($K = 1000$):



Bayesian inference:

$$\begin{aligned}
P_\beta(\{q_i\}|\{n_i\}) &= \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})} \\
P(\{n_i\}|\{q_i\}) &= \prod_{i=1}^{K} (q_i)^{n_i} \\
\langle q_i \rangle_\beta &= \frac{n_i + \beta}{N + K\beta}
\end{aligned}$$

Some common choices:

Maximum likelihood	$\beta \to 0$
Laplace's successor rule	$\beta = 1$
Krichevsky–Trofimov estimator	$\beta = 1/2$
Schurmann–Grassberger estimator	$\beta = 1/K$

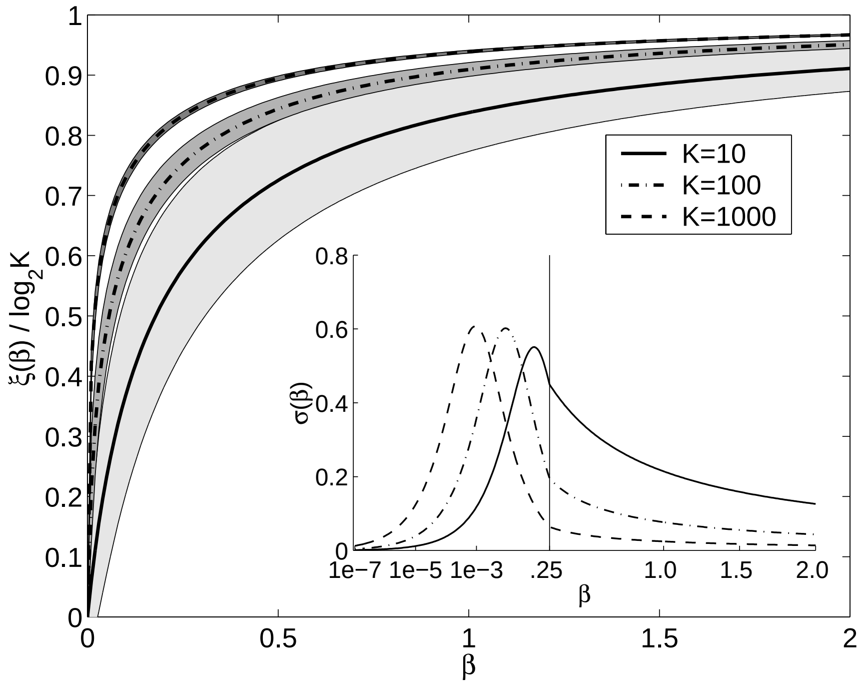
A priori expectations about the entropy:

$$\mathcal{P}_\beta(S) = \int dq_1 dq_2 \cdots dq_K \, P_\beta(\{q_i\}) \, \delta\left[S + \sum_{i=1}^{K} q_i \log_2 q_i\right]$$

The first few moments of $\mathcal{P}_\beta(S)$ are

$$\begin{aligned}
\xi(\beta) &\equiv \langle\, S[n_i = 0]\,\rangle_\beta \\
&= \psi_0(K\beta + 1) - \psi_0(\beta + 1)\,, \\
\sigma^2(\beta) &\equiv \langle\, (\delta S)^2[n_i = 0]\rangle_\beta \\
&= \frac{\beta + 1}{K\beta + 1}\,\psi_1(\beta + 1) - \psi_1(K\beta + 1) \\
\psi_m(x) &= (d/dx)^{m+1} \log_2 \Gamma(x) \text{ –the polygamma function}
\end{aligned}$$

Problem: entropy is *known a priori* for $K \gg 1$



$\xi(\beta)/\log_2 K$ and $\sigma(\beta)$ as functions of β and K; gray bands are the region of $\pm\sigma(\beta)$ around the mean. Note the transition from the logarithmic to the linear scale at $\beta = 0.25$ in the insert.

Properties:

1. Because of the phase space factors (Jacobian) of the $\{q_i\} \to S$ transformation, a priori distribution of entropy is strongly peaked.
2. The peak is narrow: $\max \sigma(\beta) = 0.61\,\text{bits} \ll \log_2 K$ at $\beta \approx 1/K$; $\sigma(\beta) \propto 1/\sqrt{K\beta}$ for $K\beta \gg 1$; $\sigma(\beta) \propto \sqrt{K\beta}$ for $K\beta \ll 1$.
3. As β varies from 0 to ∞, the peak smoothly moves from $\xi(\beta) = 0$ to $\log_2 K$. For any finite β, $\xi(\beta) = \log_2 K - O(K^0)$.

Problems:

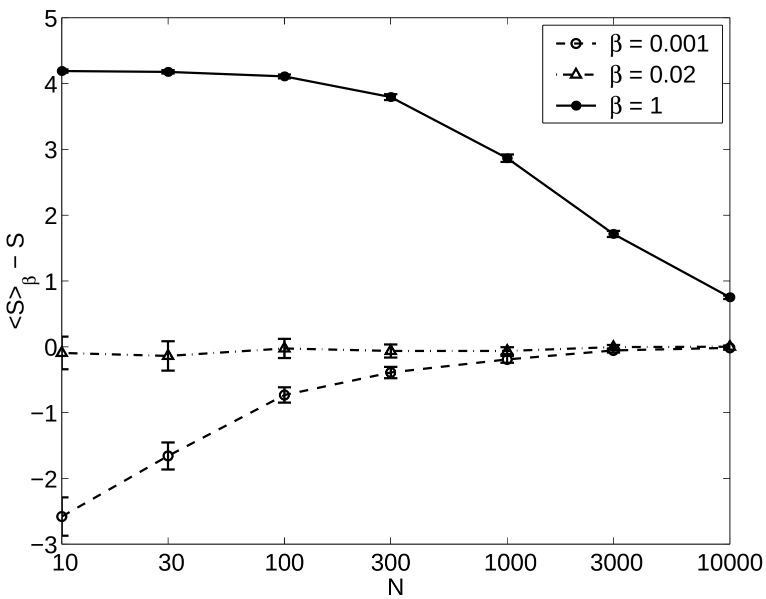
1. No a priori way to specify β.
1. Choosing β fixes allowed "shapes" of $\{q_i\}$, (cf. Panel 2) and thus defines the a priori expectation of entropy.
2. Since, for large $K\beta$, $\sigma(\beta) \sim 1/\sqrt{K\beta}$ it takes $N \sim K$ data to influence entropy estimation.
3. All common estimators (cf. Panel 3) are, therefore, bad for learning entropies.

Elaboration: problems of common estimators

Maximum likelihood: $\mathcal{P}_0(S) = \delta(S)$

1. Even $P_0(S)|_{N=1} = \delta(S)$.
2. In general, $S_{\rm ML}$ always has a downwards bias.
3. $S = S_{\rm ML} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$, $K^* = K - 1$, is an asymptotically valid correction. However, non-asymptotic choices of K^* are *ad hoc* and cannot estimate variance.

Laplace and KT: $\sigma(\beta = 1,\, 1/2) \sim 1/\sqrt{K}$



Learning the $\beta = 0.02$ distribution from Panel 2 with $\beta = 0.001, 0.02, 1$. The actual error of the estimators is plotted; the error bars are the standard deviations of the posteriors. The "wrong" estimators are very certain but nonetheless incorrect.

Schurmann–Grassberger: $\sigma(1/K) \approx 0.61$ bit.

1. Maximizes a priori entropy variance.
2. The least biased of the Dirichlet family.
3. Still strongly biased towards $S = 1/\ln 2$ bits.

Removal of the a priori bias

<u>We need</u>: such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

<u>Our options</u>:

1. $\mathcal{P}_\beta^{\text{flat}}(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}$ – difficult.
2. $\mathcal{P}(S) \sim 1 = \int \delta(S-\xi) d\xi$. Easy: $\mathcal{P}_\beta(S)$ is almost a δ-function!

<u>Solution</u>: Average over β — infinite Dirichlet mixtures

$$\mathcal{P}(\{q_i\};\beta) = \frac{1}{Z}\delta\left(1-\sum_{i=1}^{K} q_i\right)\prod_{i=1}^{K} q_i^{\beta-1}\frac{d\xi(\beta)}{d\beta}\mathcal{P}(\xi(\beta))$$

$$\widehat{S^m} = \frac{\int d\xi\, \rho(\xi,\{n_i\})\langle\, S^m[n_i]\,\rangle_{\beta(\xi)}}{\int d\xi\, \rho(\xi,[n_i])}$$

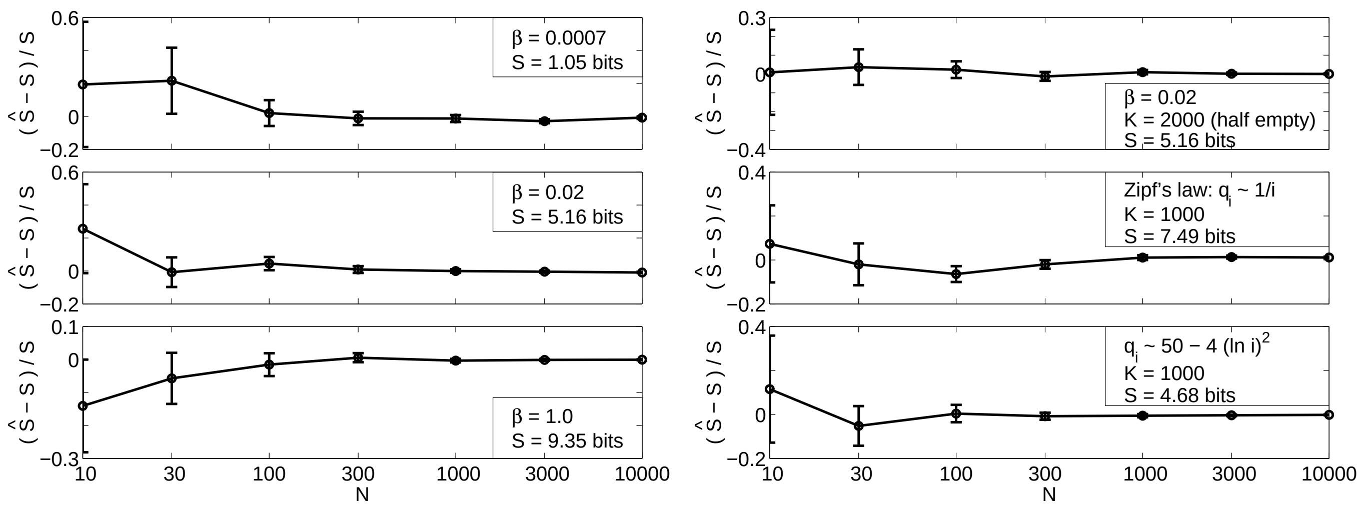
$$\rho(\xi,[n_i]) = \mathcal{P}(\xi)\frac{\Gamma(K\beta(\xi))}{\Gamma(N+K\beta(\xi))}\prod_{i=1}^{K}\frac{\Gamma(n_i+\beta(\xi))}{\Gamma(\beta(\xi))}\,.$$

<u>Notes</u>:

1. $d\xi/d\beta$ insures a priori uniformity over expected entropy.
2. $\mathcal{P}(\xi)$ embodies actual expectations about entropy.
3. Smaller β means larger allowed volume in the space of $\{q_i\}$. Thus averaging over β is Bayesian model selection (cf. Panel 1).
4. If $\rho(\xi)$ is peaked, then some $\beta(\xi)$ (model) dominates (is “selected”), and the variance of the estimator is small.

Results: unbiased estimation of entropy

Typical distributions (cf. Panel 2) | Atypical distributions



Notes:

1. Relative error $\sim 10\%$ at N as low as 30 for $K = 1000$.
2. Reliable estimation of error.
3.
   - Typical — Zipf plots like $n_i = a(\beta, N) - b(\beta)\ln i$
   - Too smooth — longer tails (e.g., Zipf's law $q_i \propto 1/i$)
   - Too rough — shorter tails (e.g., $q_i \propto 50 - 4(\ln i)^2$)
4. No bias. Possible exception: too smooth distributions.
5. Key point: learn entropies directly without finding $\{q_i\}$!

The dominant value of β saturates for typical distributions. It drifts down (towards more complex models with larger phase space) for overly rough distributions and up (towards simpler models) for too smooth cases.

N	1/2 full	Zipf	rough
units	$\cdot 10^{-2}$	$\cdot 10^{-1}$	$\cdot 10^{-3}$
10	1.7	1907	16.8
30	2.2	0.99	11.5
100	2.4	0.86	12.9
300	2.2	1.36	8.3
1000	2.1	2.24	6.4
3000	1.9	3.36	5.4
10000	2.0	4.89	4.5

V. Balasubramanian, *Neural Comp.* **9**, 349–368 (1997), `adap-org/9601001`.

W. Bialek, C. Callan, and S. Strong, *Phys. Rev. Lett.* **77**, 4693–4697 (1996), `cond-mat/9607180`.

W. Bialek, I. Nemenman, N. Tishby, *Neural Comp.* **13**, 2409-2463 (2001), `physics/0007070`.

K. Karplus, TR UCSC-CRL-95-11, UC Santa Cruz, Computer Science Department (1995).

S. Ma, *J. Stat. Phys.* **26**, 221 (1981).

D. MacKay, *Neural Comp.* **4**, 415–448 (1992).

I. Nemenman, Ph.D. Thesis, Princeton, (2000), ch. 3, `physics/0009032`.

I. Nemenman and W. Bialek, *Advances in Neural Inf. Processing Systems 13*, 287–293 (2001), `cond-mat/0009165`.

S. Panzeri and A. Treves, *Network: Comput. in Neural Syst.* **7**, 87–107 (1996).

T. Schurmann and P. Grassberger, *Chaos* **6**, 414–427 (1996).

J. Skilling, in *Maximum entropy and Bayesian methods,* J. Skilling ed. (Kluwer Academic Publ., Amsterdam, 1989), pp. 45–52.

S. Strong et al., *Phys. Rev. Lett.* **80**, 197–200 (1998), `cond-mat/9603127`.

F. Willems, Y. Shtarkov, and T. Tjalkens, *IEEE Trans. Inf. Thy.*, **41**, 653–664 (1995).

D. Wolpert and D. Wolf, *Phys. Rev. E*, **52**, 6841–6854 (1995), `comp-gas/9403001`.