Entropy estimation: coincidences, additivity, and uninformative priors

Ilya Nemenman
CCS-3/CNLS, LANL; CDDMA, NMCIAS

nsb-entropy.sf.net
H. L. Leertouwer
Characterize coding without explicit decoding

\[ S[x] = -\sum_x p(x) \log p(x), \quad x = s, \{t_i\} \]

\[ I[s, \{t_i\}] = \sum_{s,\{t_i\}} p(s, \{t_i\}) \log \frac{p(s, \{t_i\})}{p(s)p(\{t_i\})} \]

\[ = S[r] - S[r | s] \]

- Captures all dependencies (zero iff joint probabilities factorize)
- Reparameterization invariant
- Unique metric-independent measure of “how related”
- Because this it is the currency of sensory systems
Experiment design

(Strong et al., 1998)
Experiment design: probing precise spike timing

- Rate code
- Timing code
Problems

- Total of about 10-15 min of recordings (limited by stationarity of the outside world)
- At most 200 repetitions
- Stimulus correlation of 60ms: only 10000 independent samples (repeated or nonrepeated)
- Need to sample words of length 30 ms (behavioral) to 60 ms (stimulus) at resolution down to 0.2 ms (binary words of length up to ~100).

Undersampling!
Undersampling and entropy/MI estimation

Maximum likelihood estimation:

$$p_i, \quad i = 1 \ldots K$$

$$S_{ML} = - \sum_i n_i \log \frac{n_i}{N}$$

$$\langle S_{ML} \rangle \leq - \sum_i \frac{n_i}{N} \log \frac{n_i}{N} = S$$

$$p_{i}^{ML} = \frac{n_i}{N}$$

(K - # of bins)

(N - sample size)
Undersampling and entropy/MI estimation

\[ \langle S_{ML} \rangle \leq - \sum_i \frac{\langle n_i \rangle}{N} \log \frac{\langle n_i \rangle}{N} = S \]

\[ \log K \]

\[ \text{bias} \propto - \frac{2^S}{N} \gg (\text{variance})^{1/2} \propto \frac{1}{\sqrt{N}} \]

Fluctuations underestimate entropies and overestimate mutual informations.

(Need “smoothing”)
Correct smoothing possible

\[ S \leq \log N \]

Incorrect smoothing -- over- or underestimation.

Even **refractory** Poisson process at these \( T, \tau \) has over 15-20 bits of entropy!

For estimation of entropy at \( K/N<<1 \) see:
What if $S > \log N$?

But there is hope (Ma, 1981) (*capture-recapture*):

For uniform $K$-bin distribution the first coincidence occurs at

$$N_c \sim \sqrt{K} = \sqrt{2^S}$$

$$S \sim 2 \log N_c$$

Can make estimates for square-root-fewer samples!

Can this be extended to nonuniform cases?

- Assumptions needed (won’t work always)
- Estimate entropies without estimating distributions.
Generalizing Ma: What is unknown?

Binomial distribution:

\[ S = -p \log p - (1-p) \log(1-p) \]

Assume (Bayes)

uniform (no assumptions)
What is unknown?

Selection of wrong “unknown” biases the estimation.

(Even worse for large $K$.)

\[ \varepsilon = \left\langle \frac{S_{est} - S_{true}}{\delta S_{est}} \right\rangle \]
For large $K$ the problem is extreme (S known a priori)

\[
P_\beta(\{q_i\}) = \frac{1}{Z(\beta)} \delta\left(1 - \sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1}
\]

Dirichlet priors, a.k.a., adding pseudocounts (include the uniform prior, the ML prior, and others).

Inference is analytic
For large $K$ the problem is extreme (S known a priori)

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

Dirichlet priors, a.k.a., adding pseudocounts (include the uniform prior, the ML prior, and others).

Inference is analytic.

Persists for $N \sim K^{1/2}$
Uniformize on $S$

$$P_\beta(\{q_i\}, \beta) = \frac{1}{Z} \delta(1 - \sum_{i=1}^{K} q_i) \prod_{i=1}^{K} q_i^\beta \frac{dS}{d\beta}\bigg|_{N=0} P(S|_{N=0})$$

- A delta-function sliding along the a priori entropy expectation.
- This is also Bayesian model selection (small $\beta$ large phase space)
- Have error bars (dominated by posterior variance in $\beta$, not at fixed $\beta$).
Coincidence counting

\[ \Delta \equiv N - K_1; \quad K_1 = \text{#bins with } n_i \geq 1 \]

\[ \bar{S} = f(\Delta) + \text{corrections} \]

\[ \text{var } S = \frac{1}{\Delta} + \text{corrections} \]
Synthetic test (same for natural data)

Refractory Poisson, rate 0.26 spikes/ms, refractory period 1.8 ms, $T=15$ ms, discretization 0.5 ms, true entropy 13.57 bits.

- Estimator is unbiased if consistent and self-consistent.
- Always do this check.

(Nemenman et al. 2004)
NSB summary

- Posterior variance scales as $1 / \Delta$
- No bias for short-tailed distributions
- Negative bias for long-tailed distributions (strictly smaller than naïve; as for all learning, cf. Zador and DeWeese)
- Counts coincidences and works in Ma regime (if works)
- Is guaranteed correct (correct) for large $N$
- Smooth convergence: if agrees with itself for different $N$, then correct
- Allows infinite # of bins
- Not a $1/N$ series correction

(Nemenman et al. 2002-2007)
However: recall the slide

Exploring the code: long tails

This causes complications
And this slide:
Information rate at $T=25\text{ms}$

- Rate grows up to $\tau = 0.2-0.3 \text{ ms}$
- 30% more information at $\tau < 1\text{ms}$.
- $\sim 1 \text{ bit/spike at 150 spikes/s and low-entropy correlated stimulus.}$  
  Design principle?
- 0.2 ms - comparable to channel opening/closing noise and experimental noise.
How to estimate entropy for long tails?
Go to the source!

\[
S = S^{NSB}[0, else] + p_0 S_0^{NSB} + p_{else} S_{else}^{NSB}
\]
\[
= S^{NSB}[0,1, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_{else} S_{else}^{NSB}
\]
\[
= S^{NSB}[0,1,2, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_2 S_2^{NSB} + p_{else} S_{else}^{NSB}
\]
\[
= S^{NSB}[0,1,2,\ldots, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_2 S_2^{NSB} + \ldots
\]
What’s going on?

• Capture-recapture, but count perches separately from wrasses
• Within each subset, probabilities more uniform
• This is a good convergence test
• But no free lunch: more data needed, since now need coincidences in each domain
• Worst case: $2^{(S/2)}$ domains, then $N \sim 2^S$
• Usually: $N \sim 2^{(cS)}$, $c < 1$
• Conjecture: this achieves best possible performance for any distribution, whatever that performance is
Example: non NS for generality
How to estimate MI directly?

\[ I = S_1 + S_2 - S_{12} = (\text{large} + \text{large}) - (\text{twice large}) \]

\[ = \text{small, lost in noise} \]
Also recall the slide:
Precision is limited by physical noise sources

\[ T = 6 \text{ ms} \]
\[ \tau = 0.2 \text{ ms} \]
\[ 1.1 \times 10^6 \text{ ph/(s \cdot rec)} \pm 3\% \]
\[ I^+ - I^- = 0.0204 \pm 0.0108 \text{ bits} \]
\[ p = 6\% \text{ (and much smaller)} \]
Solution: MI differences (ranks)
Non NS, for generality
Is estimating MI directly worth it? NSB assumptions about MI

Judge for yourself!