# Entropy estimation: coincidences, additivity, and uninformative priors 

Ilya Nemenman<br>CCS-3/CNLS, LANL; CDDMA, NMCIAS<br>nsb-entropy.sf.net

CNS'07, Toronto
H. L. Leertouwer


## Characterize coding without explicit decoding

$$
\begin{aligned}
& S[x]=-\sum_{x} p(x) \log p(x), \quad x=s,\left\{t_{i}\right\} \\
& \begin{aligned}
I\left[s,\left\{t_{i}\right\}\right] & =\sum_{s\left\{t_{i}\right\}} p\left(s,\left\{t_{i}\right\}\right) \log \frac{p\left(s,\left\{t_{i}\right\}\right)}{p(s) p\left(\left\{t_{i}\right\}\right)} \\
& =S[r]-S[r \mid s]
\end{aligned}
\end{aligned}
$$

- 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

$$
T=4
$$



$I=S^{t}-S^{n}$


- Los Alamos

NATIONAL LABORATORY
(Strong et al., 1998)

## Experiment design: probing precise spike timing



Slide 4

## 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 $\sim 100$ ).


## Undersampling!

Slide 5

## Undersampling and entropy/MI estimation

Maximum likelihood estimation:

$$
\begin{aligned}
& \left\langle S_{M L}\right\rangle \leq-\sum_{i} \frac{\left\langle n_{i}\right\rangle}{N} \log \frac{\left\langle n_{i}\right\rangle}{N}=S
\end{aligned}
$$

NATIONAL LABORATORY

## Undersampling and entropy/MI estimation

$$
\left\langle S_{M L}\right\rangle \leq-\sum_{i} \frac{\left\langle n_{i}\right\rangle}{N} \log \frac{\left\langle n_{i}\right\rangle}{N}=S
$$

$\log K$

$$
\text { bias } \propto-\frac{2^{s}}{N} \gg(\text { variance })^{1 / 2} \propto \frac{1}{\sqrt{N}}
$$

Fluctuations underestimate entropies and overestimate mutual informations.

> (Need "smoothing")

Slide 7

## 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:
Grassberger 1989, 2003, Antos and Kontoyiannins 2002, Wyner and
Foster 2003, Batu et al. 2002, Paninski 2003, Panzeri and Treves
1996, Strong et al. 1998. Not sure how to quantify Victor 97-02.
Shlens et al is for K^}a/N<<1,a<1
```


## What if $S>\log N$ ?

But there is hope (Ma, 1981) (capture-recapture):
For uniform $K$-bin distribution the first coincidence occurs at

$$
\begin{aligned}
& N_{c} \sim \sqrt{K}=\sqrt{2^{s}} \\
& S \sim 2 \log N_{c} \longleftarrow
\end{aligned} \text { Time of first coincidence }
$$

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:

$$
\begin{aligned}
& S=-p \log p- \\
& \quad(1-p) \log (1-p)
\end{aligned}
$$



## What is unknown?



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

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

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


Slide 12

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

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

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


## Uniformize on S

$$
P_{\beta}\left(\left\{q_{i}\right\}, \beta\right)=\left.\frac{1}{Z} \delta\left(1-\sum_{i=1}^{K} q_{i}\right) \prod_{i=1}^{K} q_{i}^{\beta} \frac{d S}{d \beta}\right|_{N=0} P\left(\left.S\right|_{N=0}\right)
$$

- 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

$\Delta \equiv N-K_{1} ; \quad K_{1}=$ \#bins with $n_{i} \geq 1$
$\bar{S}=f(\Delta)+$ corrections
$\operatorname{var} S=\frac{1}{\Delta}+$ corrections

## Synthetic test (same for natural data)

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


- Estimator is unbiased if consistent and selfconsistent.
- 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



## And this slide:

 Information rate at $T=25 \mathrm{~ms}$

- Rate grows up to $\tau$ $=0.2-0.3 \mathrm{~ms}$
- $30 \%$ more information at $\tau<1 \mathrm{~ms}$.
- ~1 bit/spike at 150 spikes/s and lowentropy correlated stimulus. Design principle?
- 0.2 ms - comparable to channel opening/ closing noise and experimental noise.

NATIONAL LABORATORY

## How to estimate entropy for long tails? Go to the source!



$$
\begin{aligned}
S & =S^{N S B}[0, \text { else }]+p_{0} S_{0}^{N S B}+p_{\text {elss }} S_{\text {else }}^{N S B} \\
& =S^{N S B}[0,1, \text { else }]+p_{0} S_{0}^{N S B}+p_{1} S_{1}^{N S B}+p_{\text {elsse }} S_{\text {else }}^{N S B} \\
& =S^{N S B}[0,1,2, \text { else }]+p_{0} S_{0}^{N B}+p_{1} S_{1}^{N S B}+p_{2} S_{2}^{N B}+p_{\text {else }} S_{\text {else }}^{N S B} \\
& =S^{N S B}[0,1,2, \ldots, \text { else }]+p_{0} S_{0}^{N S B}+p_{1} S_{1}^{N S B}+p_{2} S_{2}^{N S B}+\ldots
\end{aligned}
$$

## 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: $2^{\wedge}(S / 2)$ domains, then $N \sim 2^{\wedge} S$
- Usually: $\mathrm{N} \sim 2^{\wedge}(c S), c<1$
- Conjecture: this achieves best possible performance for any distribution, whatever that performance is


## Example: non NS for generality



Slide 22
Operated by the Los Alamos National Security, LLC for the DOE/NNSA
MVEA

## How to estimate MI directly?

$$
\begin{gathered}
I=S_{1}+S_{2}-S_{12}=(\text { large }+ \text { large })-(\text { twice large }) \\
\\
=\text { small, lost in noise }
\end{gathered}
$$

## Also recall the slide: Precision is limited by physical noise sources



Slide 24

## Solution: MI differences (ranks) Non NS, for generality



## Is estimating MI directly worth it? NSB assumtions about MI



Judge for yourself!

