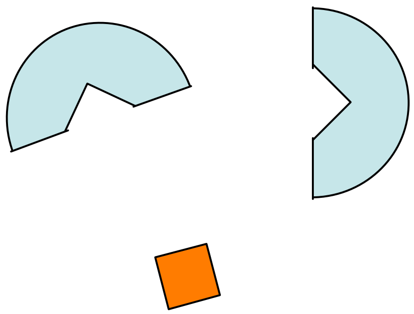
Entropy estimation: coincidences, additivity, and uninformative priors

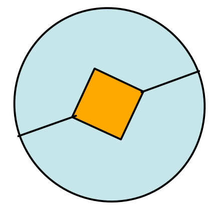
Ilya Nemenman

CCS-3/CNLS, LANL

`nsb-entropy.sf.net`

MD simulations: Does a protein bind a ligand?





$$F_1 = E_1 - TS_1 \quad \overset{\text{binding}}{<} \quad F_2 = E_2 - TS_2$$

Configuration sampling: $\{C_i, E_i\},\ i = 1 \ldots N$

$$\langle E \rangle = \sum_i \frac{E_i}{N} \quad - \quad \text{easy}$$

$$\langle S \rangle = ?$$

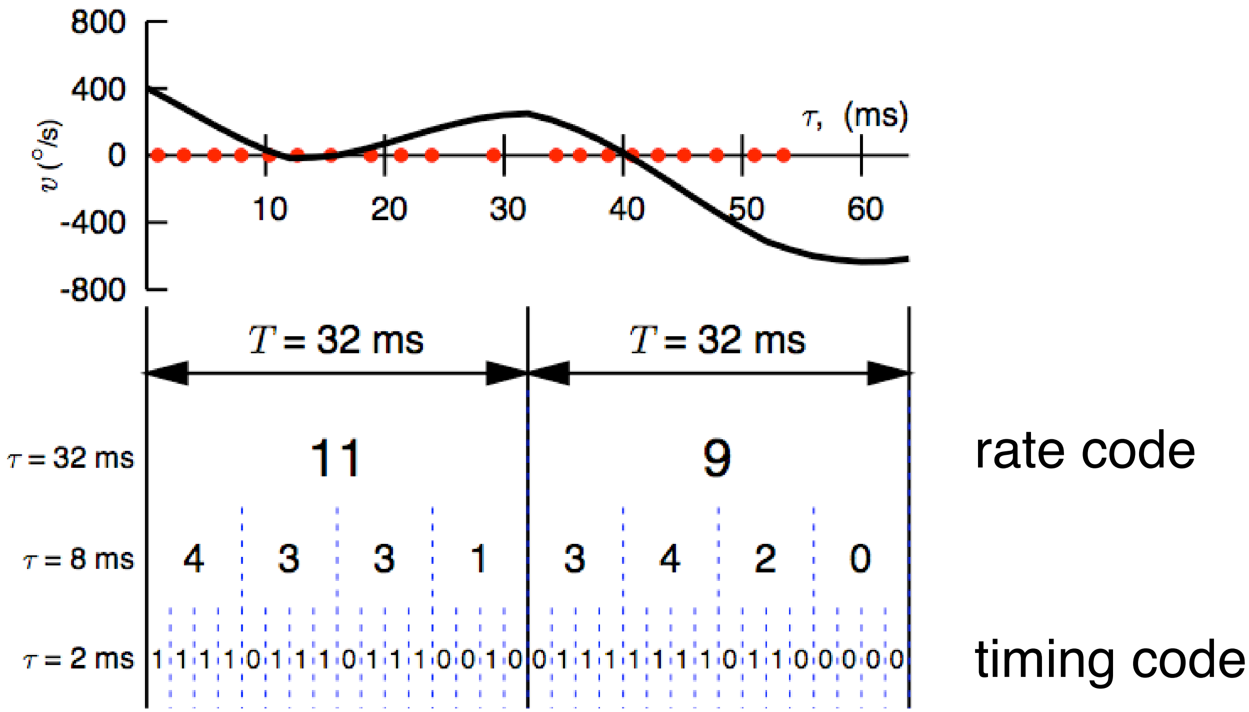
Undersampling and entropy estimation

$$\left\langle S_{ML} \right\rangle_{\{n_i\}} = \left\langle -\sum_i \frac{n_i}{N} \log \frac{n_i}{N} \right\rangle_{\{n_i\}} \leq -\sum_i p \log p = S$$

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

- Fluctuations = negative entropy bias
- Universal bias correction possible for S << log N
- *Won't work in our case (S~100s bits)*

Information content of spike trains: probing precise spike timing



T ~ 50ms, τ~ 0.2ms, L=T/τ~ 250

S up to 250 bits; 2^250 ~ 10^75 samples may be needed

(refractoriness helps, but not much)

Hope:
Coincidences and Entropy Estimation

- Catch-tag-release population sampling
  - What does a coincidence tell us?
- Recall the "birthday problem" (Ma 1981, microcanonical ensemble)

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

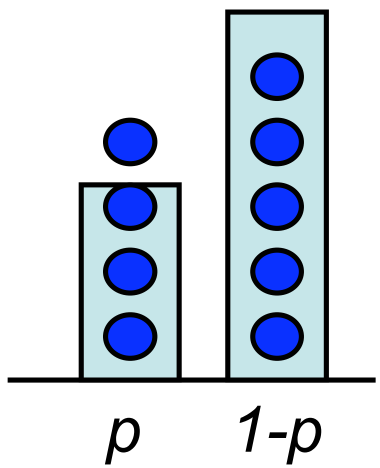
$$S \sim 2 \log N_c$$

- Can estimate entropies with square-root-fewer samples but with assumptions
- Estimate entropies directly, not distributions
- Assumptions needed (may not work always)
- What if the distribution is not uniform? (canonical ensemble)

Generalizing Ma:
What is unknown?

Binomial distribution:

$$S = -p\log p - (1-p)\log(1-p)$$

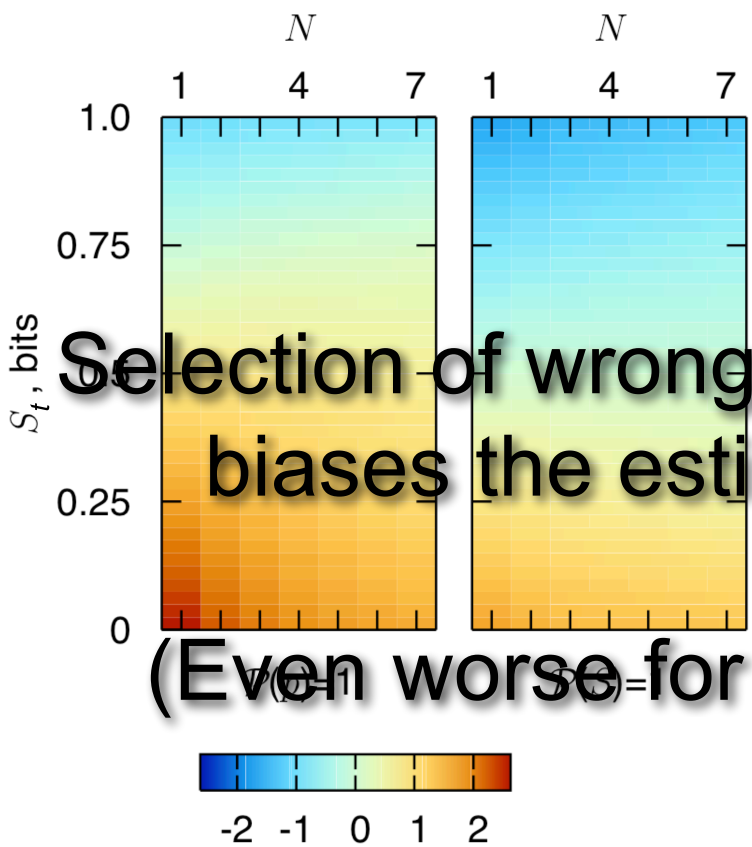


Assume (Bayes)

uniform (no assumptions)

p S

What is unknown?



$$\varepsilon = \left\langle \frac{S_{est} - S_{true}}{\delta S_{est}} \right\rangle$$

Selection of wrong “unknown” biases the estimation.

(Even worse for large *K*.)

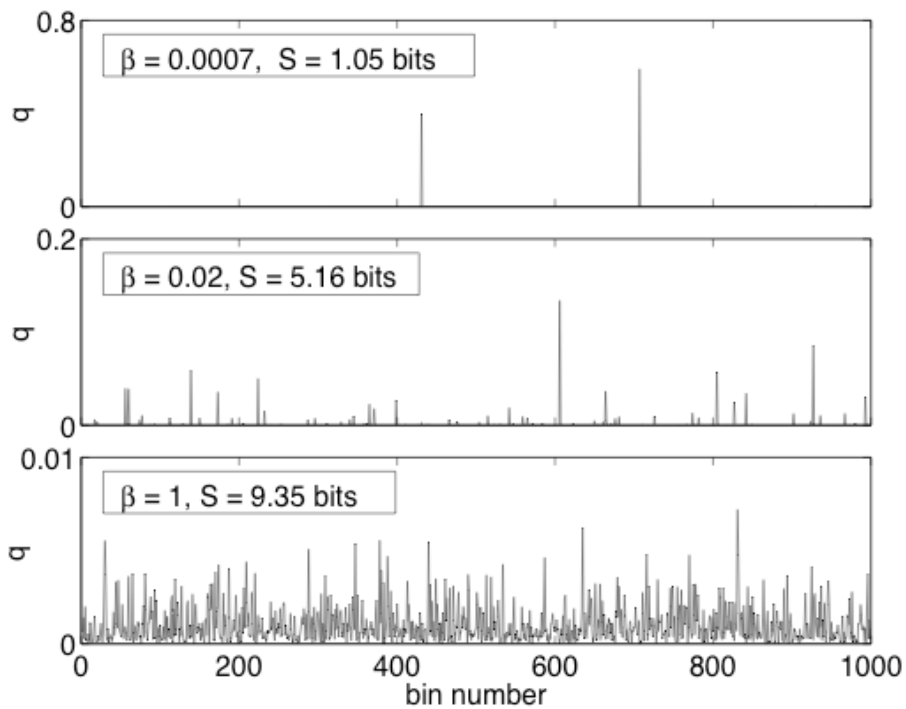
For large *K* the problem is extreme (*S* known a priori)

$$P_\beta(\{p_i\}) = \frac{1}{Z(\beta)}\,\delta\left(1-\sum\nolimits_{i=1}^{K} p_i\right)\prod_{i=1}^{K} p_i^{\beta-1}$$

$$\langle p_i \rangle = \frac{n_i+\beta}{N+K\beta}$$

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

Inference is analytic



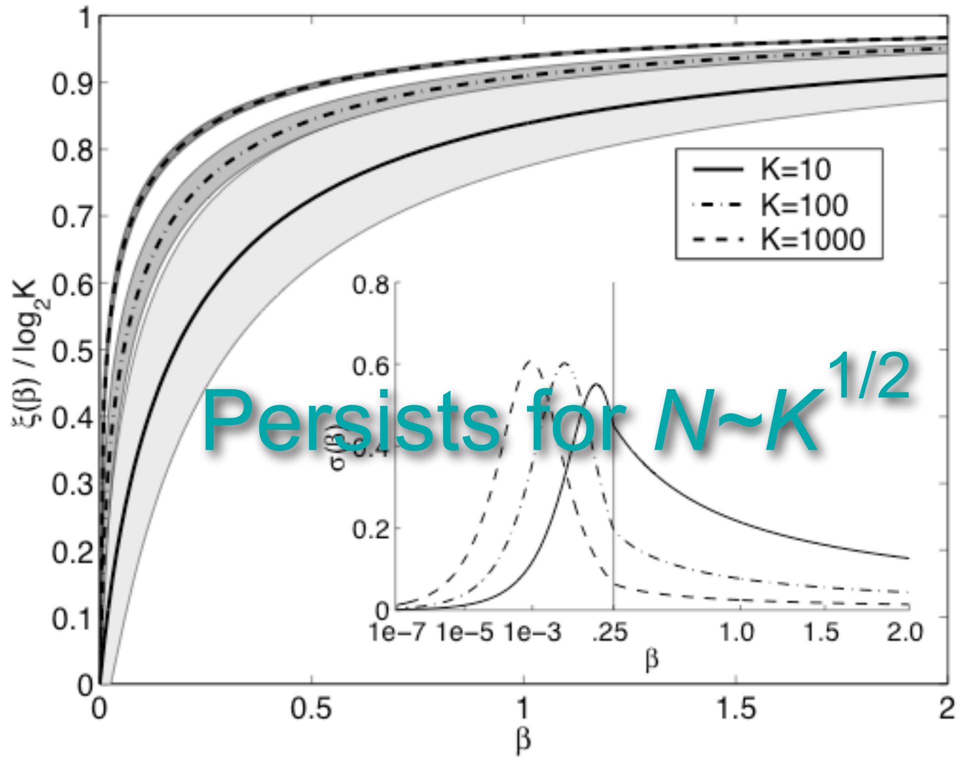
For large *K* the problem is extreme (*S* known a priori)

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

$$\langle p_i \rangle = \frac{n_i+\beta}{N+K\beta}$$

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

Inference is analytic



Uniformize on *S*

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

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

Coincidence counting

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

$$\bar{S} = f(\Delta) + \text{small corrections}$$

$$\operatorname{var} S = \frac{1}{\Delta} + \text{small corrections}$$

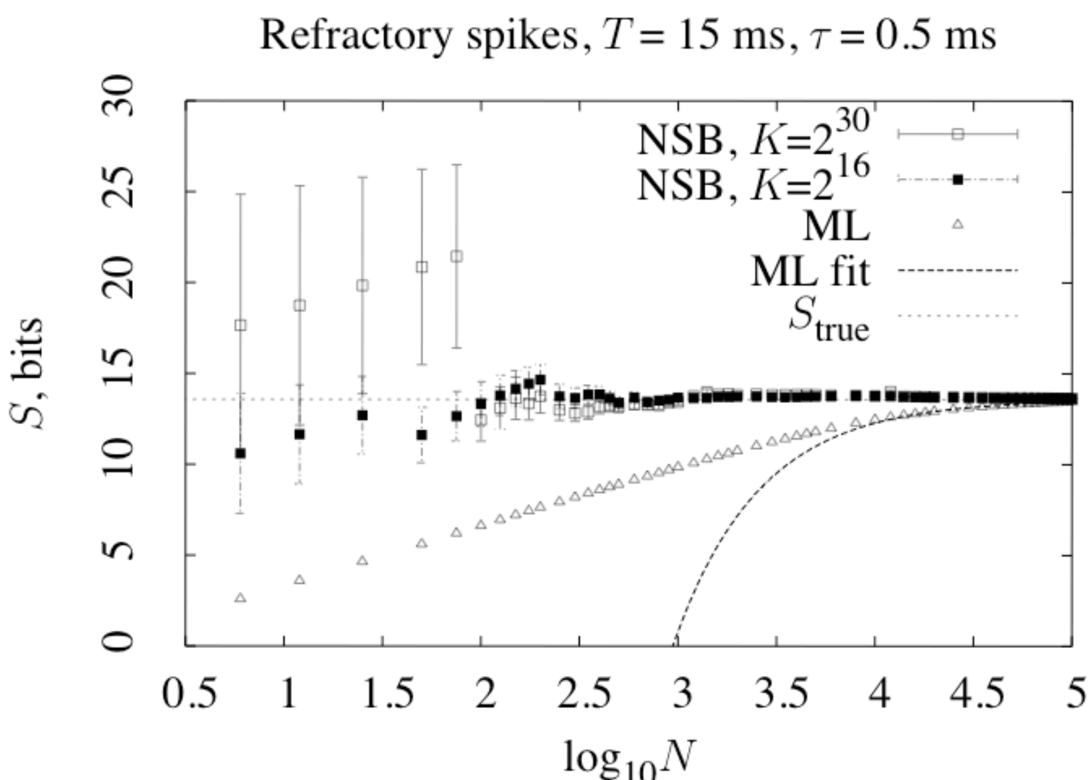
NSB summary

- 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 (consistent) for large *N*
- Smooth convergence: if agrees with itself for different *N,* then correct
- Allows infinite # of bins
- Not a 1/*N* series correction, but $1/\Delta$ expansion

(Nemenman et al. 2002-2007)

Synthetic test
(same for natural data)

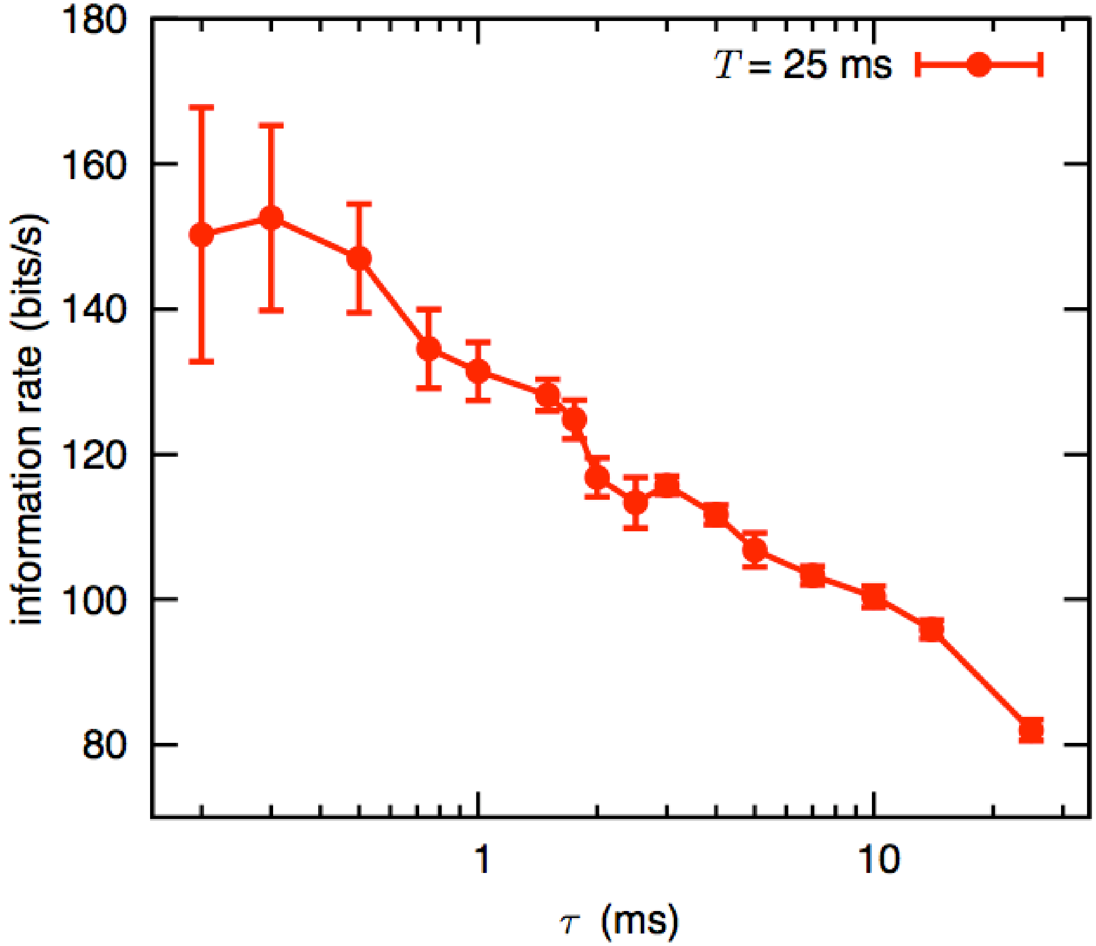
Refractory Poisson, rate 0.26 spikes/ms, refractory period 1.8 ms, *T*=15ms, discretization 0.5ms, true entropy 13.57 bits.



- NB: Estimator is unbiased if consistent and self-consistent.

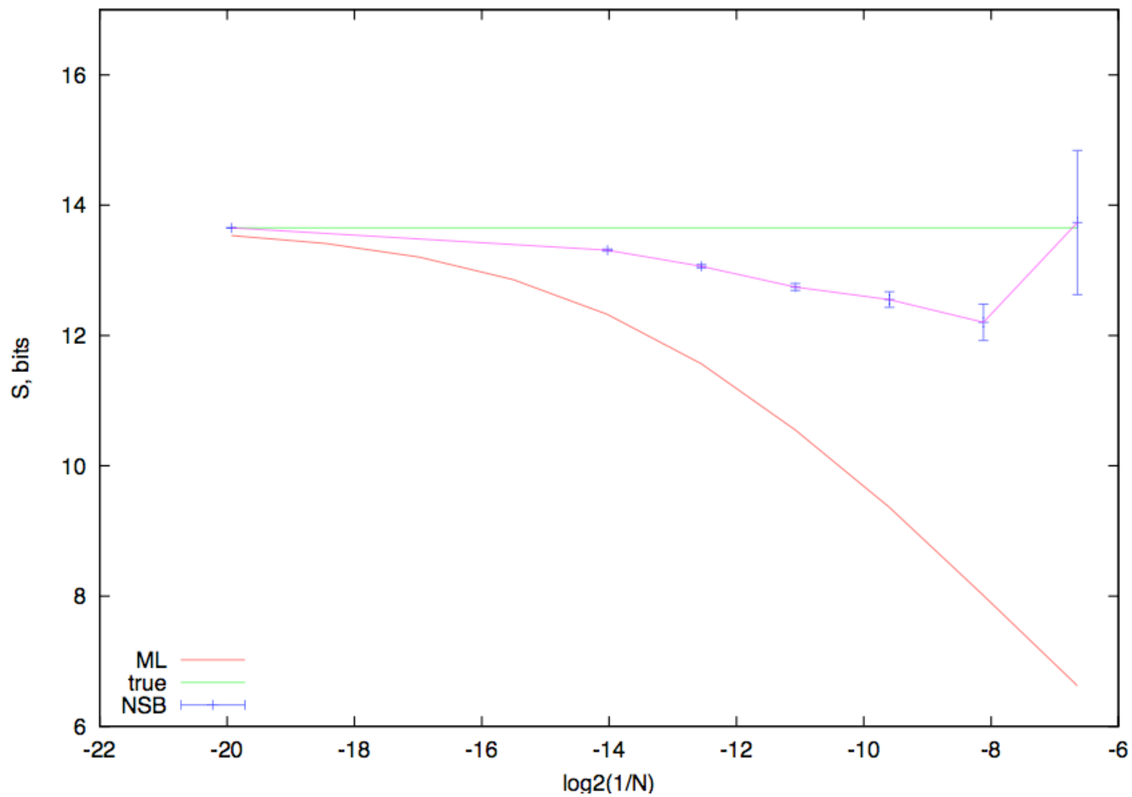
(Nemenman et al. 2004)

Neural results:
Information rate at *T*=25ms



- Rate grows up to τ =0.2-0.3 ms
- 30% more information at τ<1ms.
- ~1 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.

However:
Long tails for lattice proteins

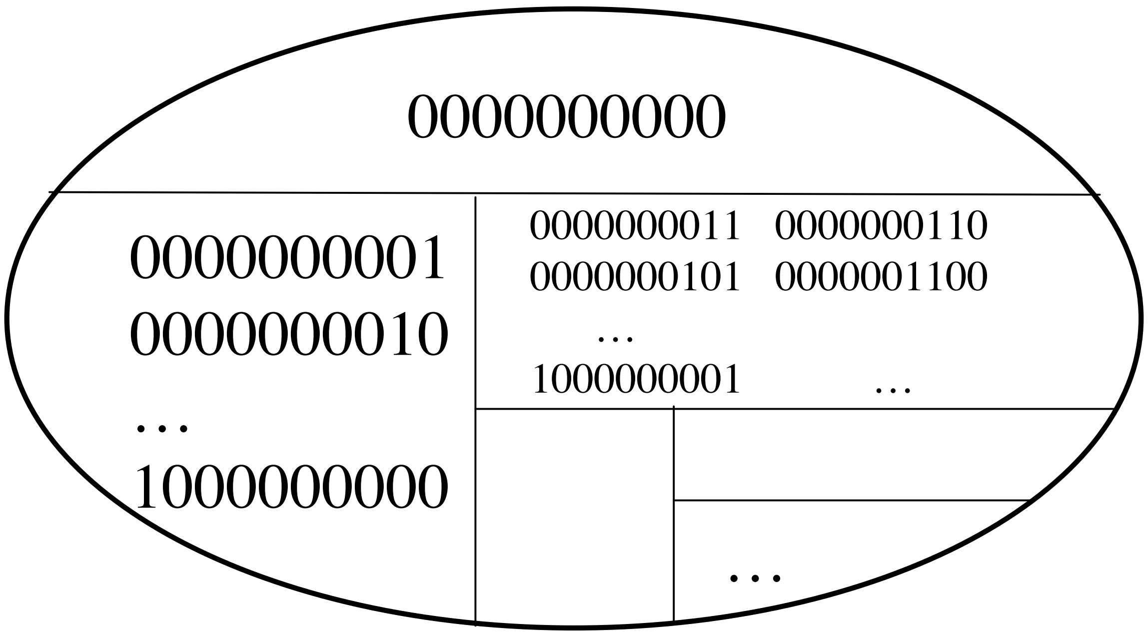


2^7 samples for NSB=
2^14 samples for ML

NSB fails!
(though always better than ML)

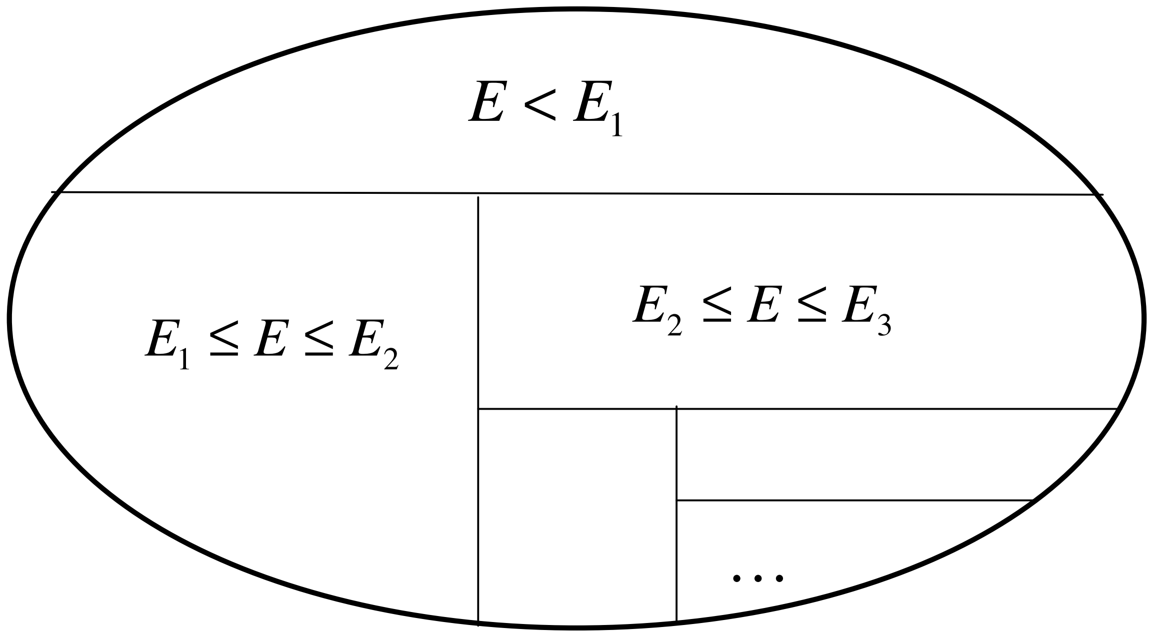
chiral, cubic, 4x4x4, 32 residues

How to estimate entropy for long tails?
Go to the source: entropy is additive!



$$
\begin{aligned}
S &= S^{NSB}[0, else] + p_0 S_0^{NSB} + p_{else} S_{else}^{NSB} \\
&= S^{NSB}[0,1, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_{else} S_{else}^{NSB} \\
&= S^{NSB}[0,1,2, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_2 S_2^{NSB} + p_{else} S_{else}^{NSB} \\
&= S^{NSB}[0,1,2,\ldots, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_2 S_2^{NSB} + \ldots
\end{aligned}
$$

How to estimate entropy for long tails? Go to the source: entropy is additive!

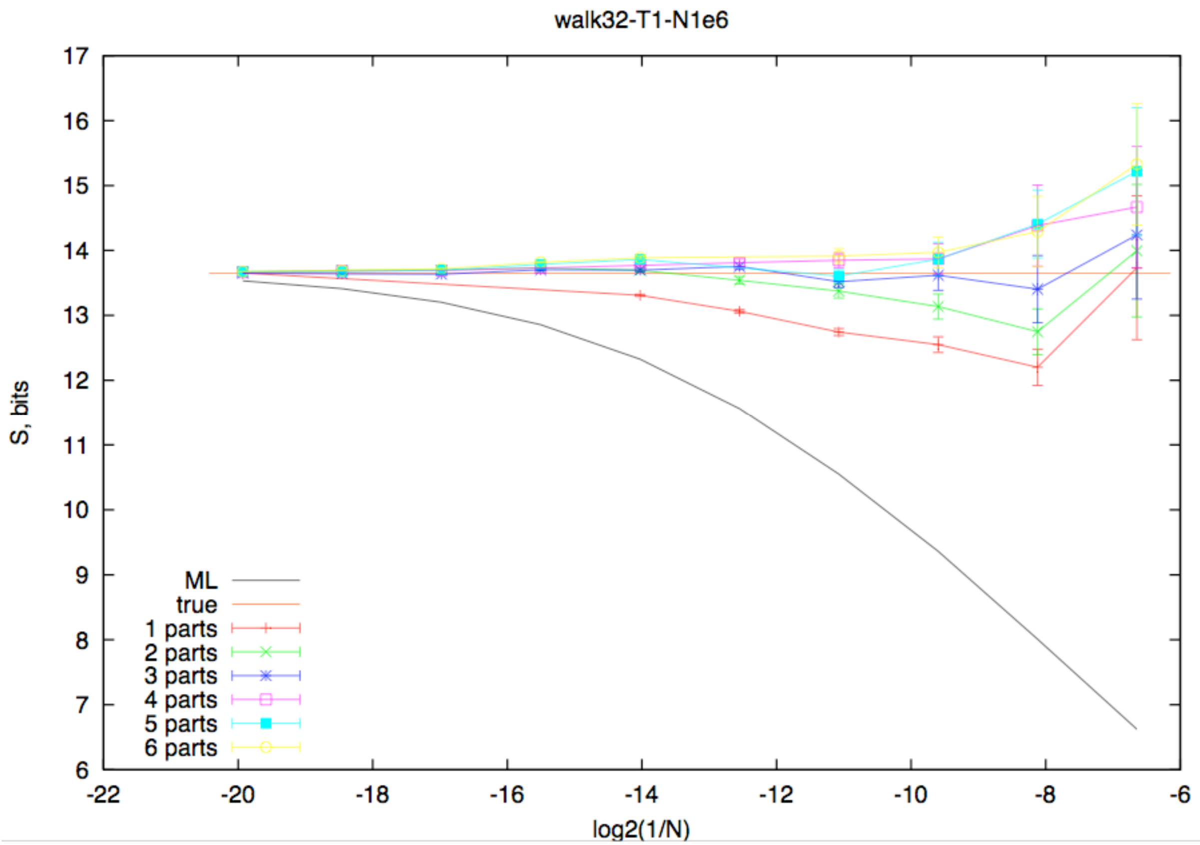


$$
\begin{aligned}
S &= S^{NSB}[0, else] + p_0 S_0^{NSB} + p_{else} S_{else}^{NSB} \\
&= S^{NSB}[0,1, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_{else} S_{else}^{NSB} \\
&= S^{NSB}[0,1,2, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_2 S_2^{NSB} + p_{else} S_{else}^{NSB} \\
&= S^{NSB}[0,1,2,\dots, else] + p_0 S_0^{NSB} + p_1 S_1^{NSB} + p_2 S_2^{NSB} + \dots
\end{aligned}
$$

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^(*S/2)* domains, then *N*~2^*S*
- Usually: N~2^(*cS*), *c*<1
- Conjecture: this achieves best possible performance for any distribution, whatever that performance is

Lattice protein entropies: It works!



There's a partition # when the estimator is unbiased!

Summary

- NSB estimator
  - Choose the right unknown
  - *or* Count coincidences
  - *or* Do model selection
- Not universal
  - But using additivity comes quite close!
- Current applications
  - Neuroscience
  - Protein structure
  - Linguistics