Estimating entropy and information in biological data

Ilya Nemenman

William Bialek, Fariel Shafee, Rob de Ruyter van Steveninck

(UCSB, Princeton University, Indiana University)

http://arxiv.org/abs/physics/0306063
http://arxiv.org/abs/physics/0207009
http://arxiv.org/abs/physics/0108025
http://arxiv.org/abs/physics/0103088

Talk outline

Problem setup Why bother?

Talk outline

Problem setup Why bother?

Developing intuition Why hard?

Talk outline

Problem setup Why bother?

Developing intuition Why hard?

The method An idea, analysis, asymptotics.

Talk outline

Problem setup Why bother?

Developing intuition Why hard?

The method An idea, analysis, asymptotics.

Applications Synthetic and natural data.

Why do we need to estimate entropies?

- information content of (symbolic) sequences

Why do we need to estimate entropies?

- information content of (symbolic) sequences
  - biology
    - information in spike trains
    - information content in molecular cell signals
    - genomic data
    - mutual information based gene expression clustering

Why do we need to estimate entropies?

- information content of (symbolic) sequences
  - biology
    - information in spike trains
    - information content in molecular cell signals
    - genomic data
    - mutual information based gene expression clustering
  - linguistics
    - comparative (historical) language analysis
    - origins and authorship of texts
    - cryptography

Why do we need to estimate entropies?

- information content of (symbolic) sequences
  - biology
    - information in spike trains
    - information content in molecular cell signals
    - genomic data
    - mutual information based gene expression clustering
  - linguistics
    - comparative (historical) language analysis
    - origins and authorship of texts
    - cryptography
  - financial data and other prediction games (Cover)

Why do we need to estimate entropies?

- information content of (symbolic) sequences
  - biology
    * information in spike trains
    * information content in molecular cell signals
    * genomic data
    * mutual information based gene expression clustering
  - linguistics
    * comparative (historical) language analysis
    * origins and authorship of texts
    * cryptography
  - financial data and other prediction games (Cover)
- dimensions of strange attractors (Grassberger et al.)
- complexity of dynamics

Genomic applications

$$\begin{array}{c|c|ccc}
\text{gccta} & \overbrace{\text{accGt}}^{N} & \text{ggtcca} & \overbrace{\text{tatata}}^{M} & \text{aggaa} \\
\text{gccta} & \text{accAt} & \text{ggtcca} & \text{tatata} & \text{tggac} \\
\text{accta} & \text{accAt} & \underbrace{\text{ggtcga}}_{D} & \text{tatata} & \text{aggac}
\end{array}$$

Genomic applications

$$\begin{array}{c|c|ccc}
 & \overbrace{\hphantom{\text{accGt}}}^{N} & & \overbrace{\hphantom{\text{tatata}}}^{M} & \\
\text{gccta} & \text{accGt} & \text{ggtcca} & \text{tatata} & \text{aggaa} \\
\text{gccta} & \text{accAt} & \text{ggtcca} & \text{tatata} & \text{tggac} \\
\text{accta} & \text{accAt} & \text{ggtcga} & \text{tatata} & \text{aggac} \\
 & & \underbrace{\hphantom{\text{ggtcca}}}_{D} & &
\end{array}$$

- along a genome
  - search for structures, possibly motifs, (overrepresented sequences) $I(M, N; D)$
  - finding conserved elements: sequences with small predictive entropies
  - running IB to extract predictive sequences

Genomic applications

$$\begin{array}{c|c|ccc} & \overbrace{\phantom{\text{accGt}}}^{N} & & \overbrace{\phantom{\text{tatata}}}^{M} & \\ \text{gccta} & \text{accGt} & \text{ggtcca} & \text{tatata} & \text{aggaa} \\ \text{gccta} & \text{accAt} & \text{ggtcca} & \text{tatata} & \text{tggac} \\ \text{accta} & \text{accAt} & \text{ggtcga} & \text{tatata} & \text{aggac} \\ & & \underbrace{\phantom{\text{ggtcga}}}_{D} & & \end{array}$$

- along a genome
  - search for structures, possibly motifs, (overrepresented sequences) $I(M, N; D)$
  - finding conserved elements: sequences with small predictive entropies
  - running IB to extract predictive sequences
- across genomes
  - estimating mutation rates
  - calculating divergence times and building phylogenetic trees
  - identifying haplotypes

Genomic applications

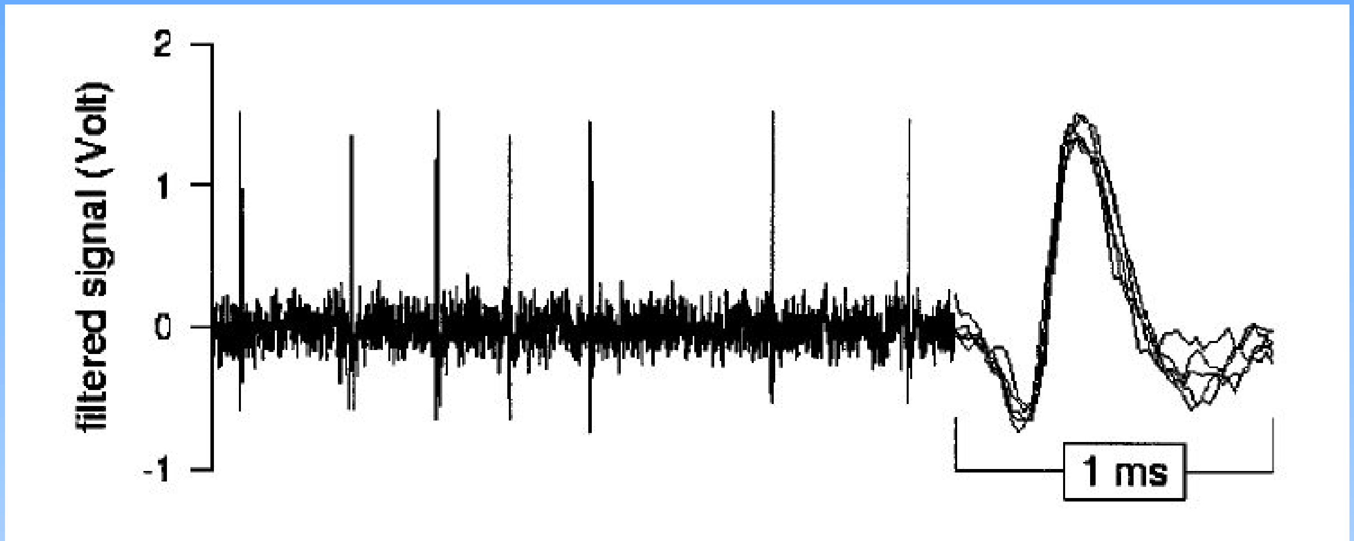
$$\begin{array}{lllll} \text{gccta} & \overbrace{\text{accGt}}^{N} & \text{ggtcca} & \overbrace{\text{tatata}}^{M} & \text{aggaa} \\ \text{gccta} & \text{accAt} & \text{ggtcca} & \text{tatata} & \text{tggac} \\ \text{accta} & \text{accAt} & \underbrace{\text{ggtcga}}_{D} & \text{tatata} & \text{aggac} \end{array}$$

- length $10^6 \dots 10^9$
- N, M, D up to 20
- < 100 repeats

Severe undersampling along.

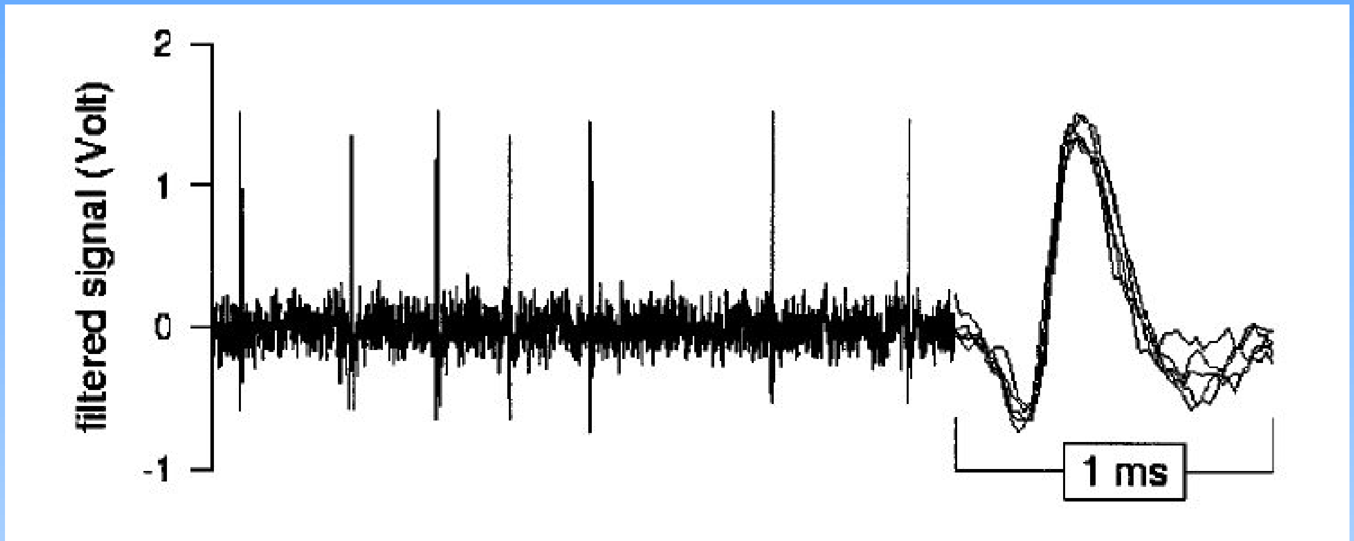
- along a genome
  - search for structures, possibly motifs, (overrepresented sequences) $I(M, N; D)$
  - finding conserved elements: sequences with small predictive entropies
  - running IB to extract predictive sequences
- across genomes
  - estimating mutation rates
  - calculating divergence times and building phylogenetic trees
  - identifying haplotypes

Neurophysiological applications



(Strong et al., 1998)

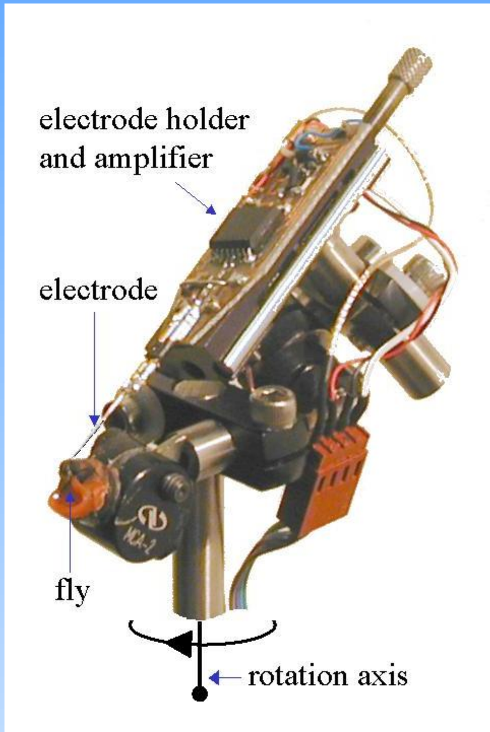
Neurophysiological applications



(Strong et al., 1998)

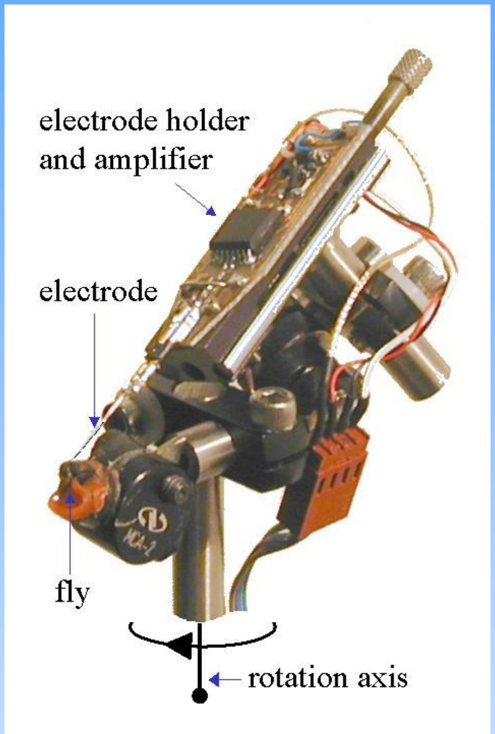
Neurons communicate by stereotypical pulses (spikes). Information is transmitted by spike rates and (possibly) precise positions of the spikes.

Experimental setup

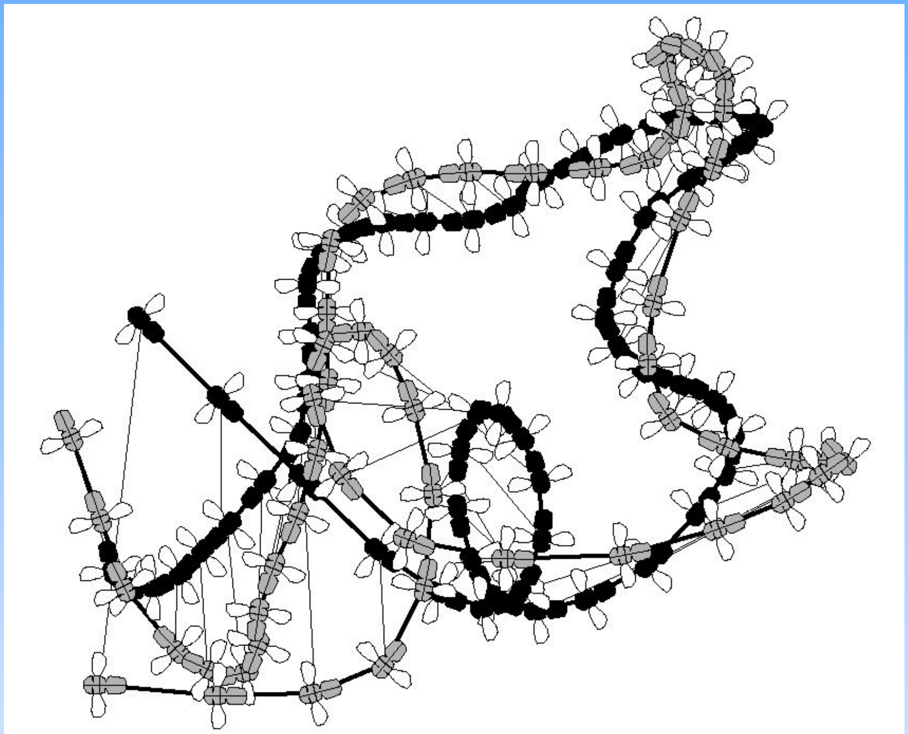


(Lewen, Bialek, and de Ruyter van Steveninck, 2001)

Experimental setup

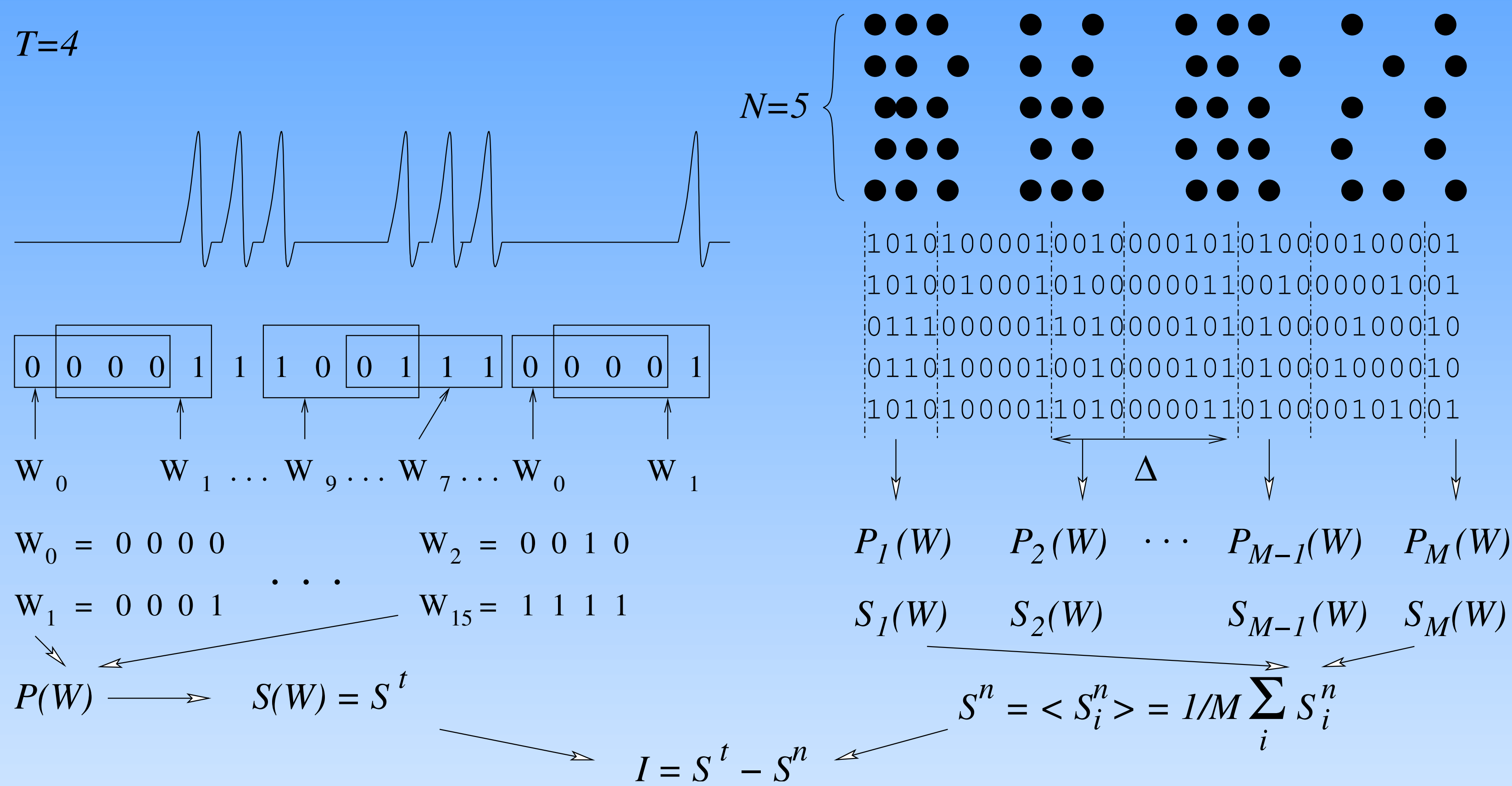


(Lewen, Bialek, and de Ruyter van Steveninck, 2001)



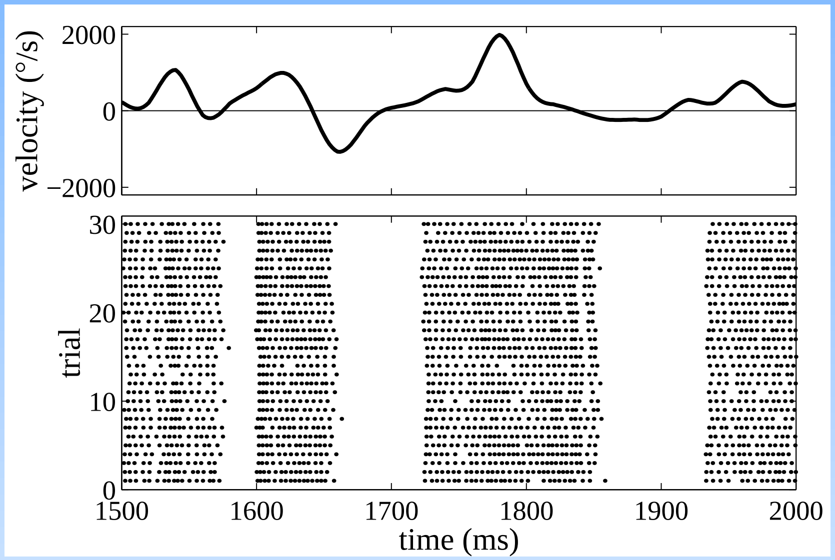
(Bialek and de Ruyter van Steveninck, 2002; Land and Collett 1974)

Estimating information rate in spike trains



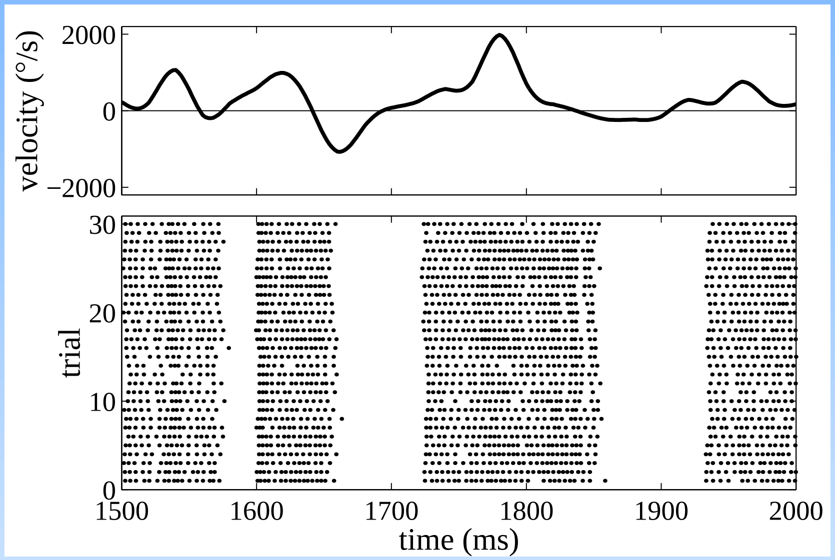
Recordings and problems

100–200 repeats of 5–10 s roller coasters rides



Recordings and problems

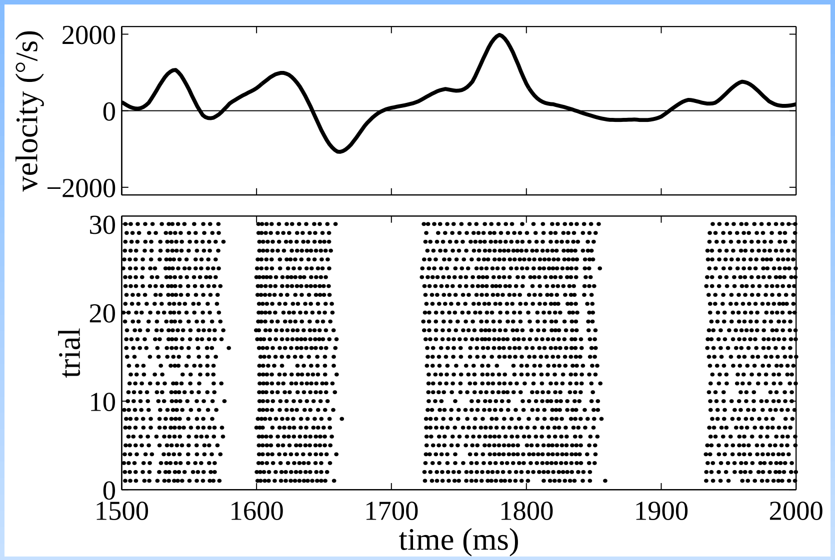
100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.

Recordings and problems

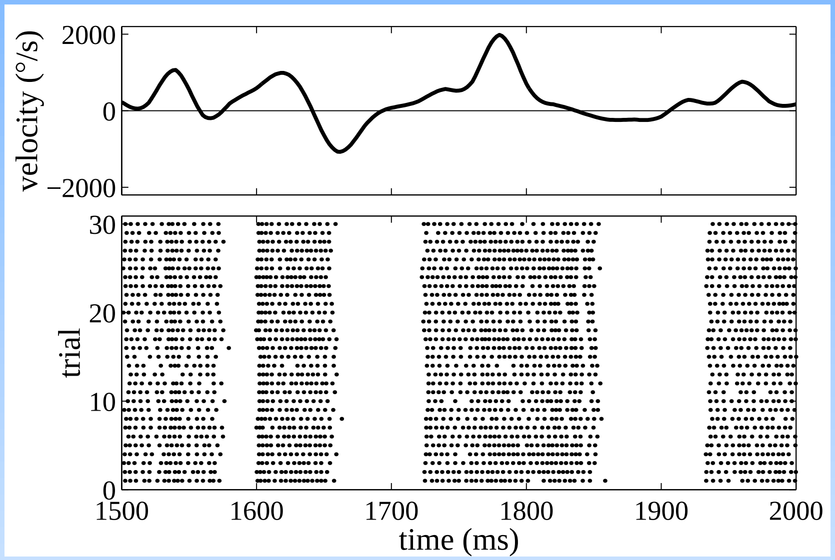
100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.
2. Need to take $\tau \to 0$ and see limiting behavior.

Recordings and problems

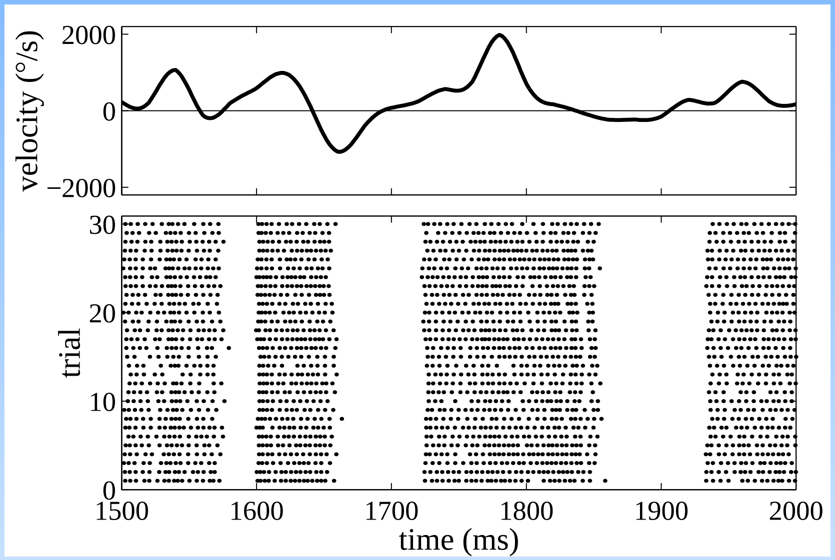
100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.
2. Need to take $\tau \to 0$ and see limiting behavior.
3. Interested in analyzing $\tau \leq 1$ms.

Recordings and problems

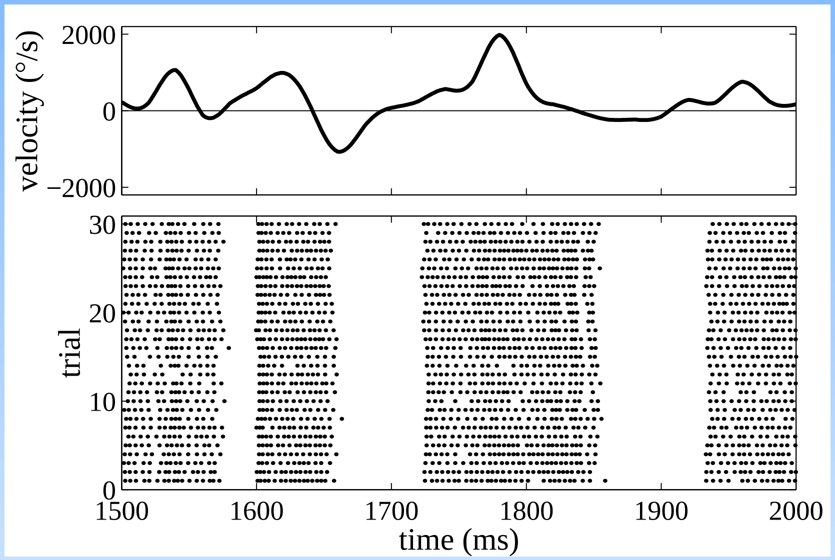
100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.
2. Need to take $\tau \to 0$ and see limiting behavior.
3. Interested in analyzing $\tau \leq 1$ms.
4. Need to have $\Delta \approx 100$ms due to natural stimulus correlations.

Recordings and problems

100–200 repeats of 5–10 s roller coasters rides



1. Need to take $T \to \infty$, $T > 30$ms for behavioral resolution.
2. Need to take $\tau \to 0$ and see limiting behavior.
3. Interested in analyzing $\tau \leq 1$ms.
4. Need to have $\Delta \approx 100$ms due to natural stimulus correlations.

Need to estimate entropies of words of length ~ 40 from < 200 samples. Undersampled!

Why is this a difficult problem?

An asymptotically $(K/N \to 0)$ easy problem.

But for $K \gg N$?

Why is this a difficult problem?

An asymptotically ($K/N \to 0$) easy problem.

But for $K \gg N$?

$$\lim_{p\to 0} \frac{p \log p}{p} = \infty$$

Why is this a difficult problem?

An asymptotically $(K/N \to 0)$ easy problem.

But for $K \gg N$?

$$\begin{aligned} \lim_{p\to 0} \frac{p \log p}{p} &= \infty \\ S_{\mathrm{ML}} &\equiv -\hat{p}\log\hat{p} - (1-\hat{p})\log(1-\hat{p}) \text{ is convex} \\ &\Longrightarrow E\, S_{\mathrm{ML}} < S(E\,\hat{p}) = S(p) \end{aligned}$$

- events of negligible probability may have large entropy (Batu et al., 2002)
- small errors in $p \Longrightarrow$ large errors in S
- unknown negative bias, variance is much smaller

- events of negligible probability may have large entropy (Batu et al., 2002)
- small errors in $p \Longrightarrow$ large errors in S
- unknown negative bias, variance is much smaller
- no finite variance unbiased entropy estimators; huge variance, small bias, but nonmonotonic is possible (Grassberger, 2003)
- no universally consistent multiplicative entropy estimator for $N/K \to 0$, $K \to \infty$ (Batu et al., 2002)
- universal consistent entropy estimation is possible only for $K/N \to \mathrm{const}$, $K \to \infty$ (Paninski, 2003)

How do others do?

For $K \gg N$:

- LZ (string matching and plug-in) (Antos and Kontoyiannis, 2002; Wyner and Foster, 2003)
  - universally consistent (under mild conditions)
  - no universal rate–of–convergence results exist for either
  - for any such universal estimator, there is always a bad distribution such that bias $\sim 1/\log N$

How do others do?

For $K \gg N$:

- LZ (string matching and plug-in) (Antos and Kontoyiannis, 2002; Wyner and Foster, 2003)
  - universally consistent (under mild conditions)
  - no universal rate–of–convergence results exist for either
  - for any such universal estimator, there is always a bad distribution such that bias $\sim 1/\log N$
- correcting for bias as a power series in $2^S/N$
  - replica–averaging over samples (Panzeri and Treves, 1996)
  - least bias $+$ variance (Paninski, 2003; Grassberger, 2003)
  - empirical evaluation of bias (Strong et al., 1998); so far the best
  - ALL WORK FOR $2^S \ll N \ll K$

The hope

Ma's (1981) argument, the birthday problem.

For uniform K–bin distribution: for $N_c \sim \sqrt{K}$, probability of coincidences ~ 1.

$$S = \log K \approx \log N_c^2 = 2 \log N_c$$

Works in nonasymptotic regime $N \sim 2^{1/2S}$. Better than it should!

$\delta S \sim 1$, but this is all we often need.

Extensions?

For Ma–type ideas to work for nonuniform cases

- forget universality, make assumptions about distributions
- do not learn distributions, learn entropies
- equate smoothness and long tails as high entropy (rapidly decaying Zipf plot)

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

$\{q_i\}$, $i = 1 \dots K$:

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

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

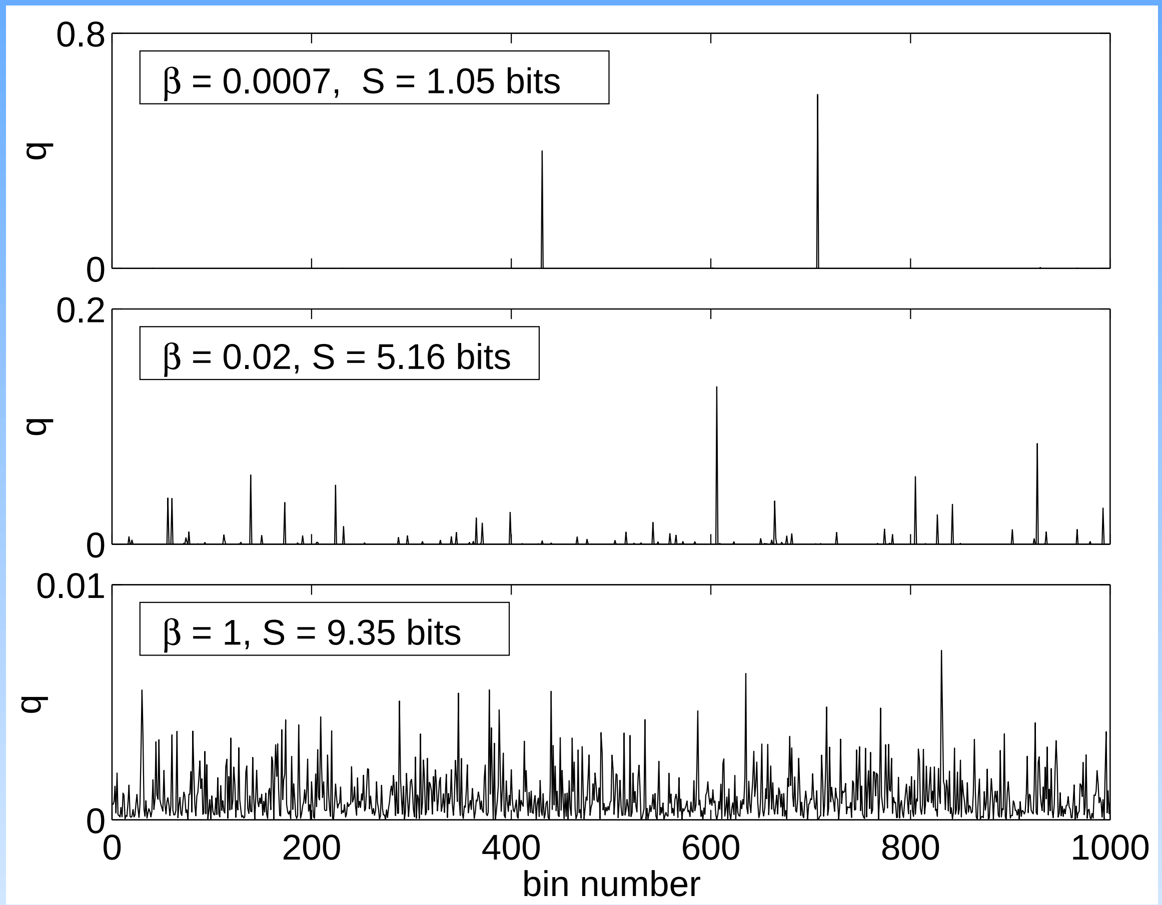
$\{q_i\}$, $i = 1 \dots K$:

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

<u>Some common choices</u>:

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

Typical distributions for $K = 1000, S \approx 9.97$



Typical rank–ordered plots

$$
\begin{aligned}
q_i &\approx 1 - \left[\frac{\beta B(\beta, \kappa-\beta)(K-1)\, i}{K}\right]^{1/(\kappa-\beta)}, \quad i \ll K\,, \\
q_i &\approx \left[\frac{\beta B(\beta, \kappa-\beta)(K-i+1)}{K}\right]^{1/\beta}, \quad K-i+1 \ll K
\end{aligned}
$$

Usually only the first regime is observed.

Gets to zero at finite i.

Faster decaying – too rough.

Slower decaying – too smooth.

Bayesian inference with Dirichlet priors

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

Bayesian inference with Dirichlet priors

$$P_\beta(\{q_i\}|\{n_i\}) = \frac{P(\{n_i\}|\{q_i\})\mathcal{P}_\beta(\{q_i\})}{P_\beta(\{n_i\})}$$

$$P(\{n_i\}|\{q_i\}) = \prod_{i=1}^{K}(q_i)^{n_i}$$

$$\langle q_i\rangle_\beta = \frac{n_i+\beta}{N+K\beta}$$

Equal pseudocounts added to each bin.

Bayesian inference with Dirichlet priors

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

Equal pseudocounts added to each bin.

Larger β means less sensitivity to data, thus more smoothing.

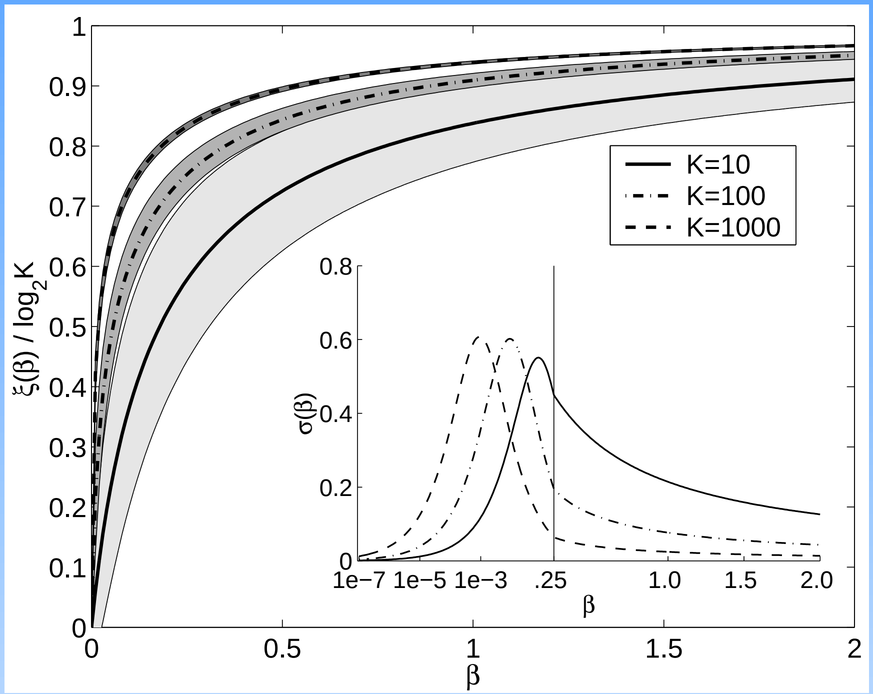
A problem: A priori entropy expectation

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

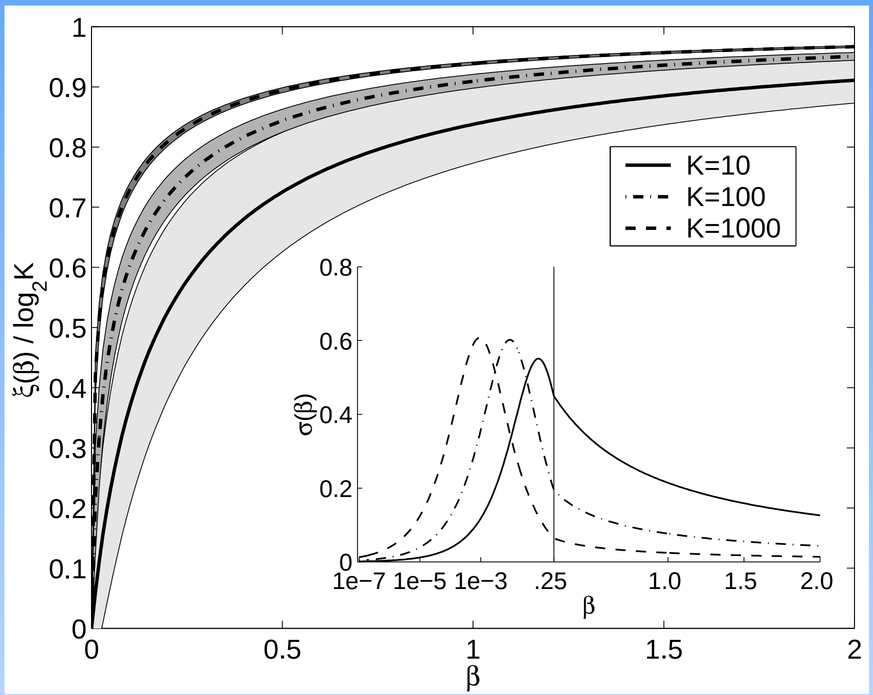
A problem: A priori entropy expectation

$$
\begin{aligned}
\mathcal{P}_\beta(S) &= \int dq_1 dq_2 \cdots dq_K \, P_\beta(\{q_i\}) \, \delta \left[S + \sum_{i=1}^{K} q_i \log_2 q_i \right] \\
\xi(\beta) &\equiv \langle \, S[n_i = 0] \, \rangle_\beta \\
&= \psi_0(K\beta + 1) - \psi_0(\beta + 1) \, , \\
\sigma^2(\beta) &\equiv \langle \, (\delta S)^2 [n_i = 0] \rangle_\beta \\
&= \frac{\beta + 1}{K\beta + 1} \psi_1(\beta + 1) - \psi_1(K\beta + 1) \\
\psi_m(x) &= (d/dx)^{m+1} \log_2 \Gamma(x) \text{ –the polygamma function}
\end{aligned}
$$

The problem: Analysis

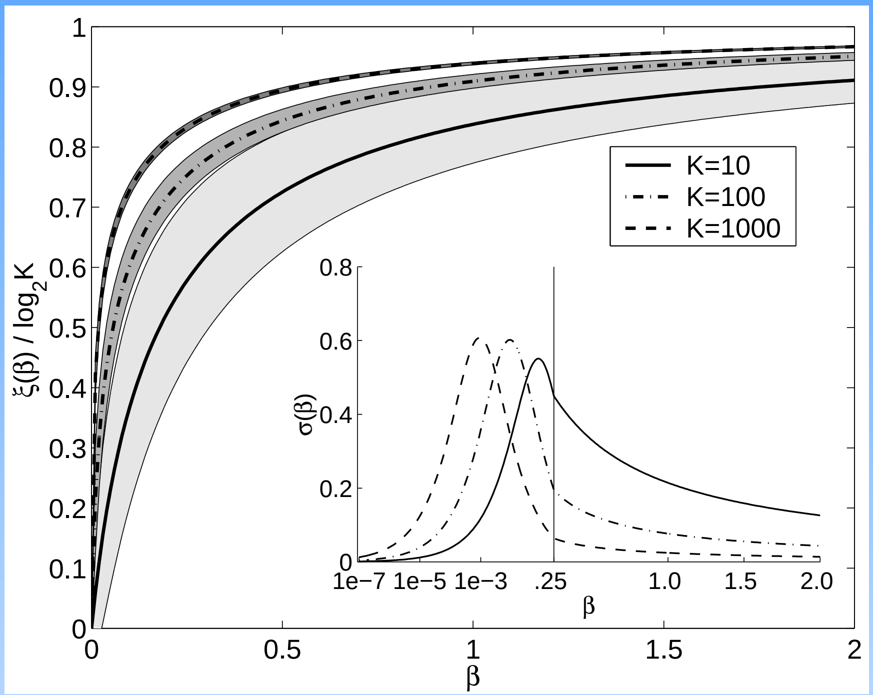


The problem: Analysis



- Because of the Jacobian of $\{q_i\} \to S$, a priori distribution of entropy is strongly peaked.
- *Narrow* peak: $\sigma(\beta) \propto 1/\sqrt{K\beta}$, $\max \sigma(\beta) = 0.61\,\mathrm{bits}$.
- As β varies from 0 to ∞, the peak smoothly moves from 0 to $\log_2 K$. For $\beta \sim 1$, $\xi(\beta) = \log_2 K - O(K^0)$.

The problem: Analysis



- Because of the Jacobian of $\{q_i\} \to S$, a priori distribution of entropy is strongly peaked.
- *Narrow* peak: $\sigma(\beta) \propto 1/\sqrt{K\beta}$, $\max \sigma(\beta) = 0.61\,\mathrm{bits}$.
- As β varies from 0 to ∞, the peak smoothly moves from 0 to $\log_2 K$. For $\beta \sim 1$, $\xi(\beta) = \log_2 K - O(K^0)$.

- No a priori way to specify β.
- Choosing β fixes allowed "shapes" of $\{q_i\}$, and defines the a priori expectation of entropy.
- Such expectation dominates data until $N \gg K\beta$.
- All common estimators are, therefore, bad for learning entropies.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Our options:

1. $\mathcal{P}_\beta^{\mathrm{flat}}(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}$.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Our options:

1. $\mathcal{P}^{\mathrm{flat}}_\beta(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}$. Difficult.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Our options:

1. $\mathcal{P}_\beta^{\rm flat}(\{q_i\}) = \frac{\mathcal{P}_\beta(\{q_i\})}{\mathcal{P}_\beta(S[q_i])}$. Difficult.
2. $\mathcal{P}(S) \sim 1 = \int \delta(S-\xi) d\xi$.

Removing the entropy bias at the source

Need such $\mathcal{P}(\{q_i\})$ that $\mathcal{P}(S[q_i])$ is (almost) uniform.

Our options:

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

Solution

Average over β — infinite Dirichlet mixtures.

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

Solution

Average over β — infinite Dirichlet mixtures.

$\beta \to \xi$ Jacobian entropy prior

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

Solution

Average over β — infinite Dirichlet mixtures.

$\beta \to \xi$ Jacobian entropy prior

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

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

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

Solution

Average over β — infinite Dirichlet mixtures.

$$\mathcal{P}(\{q_i\};\beta) = \frac{1}{Z}\,\delta\left(1 - \sum_{i=1}^{K} q_i\right) \prod_{i=1}^{K} q_i^{\beta-1} \qquad \underbrace{\frac{d\xi(\beta)}{d\beta}}_{\beta \to \xi \text{ Jacobian}} \qquad \underbrace{\mathcal{P}(\xi(\beta))}_{\text{entropy prior}}$$

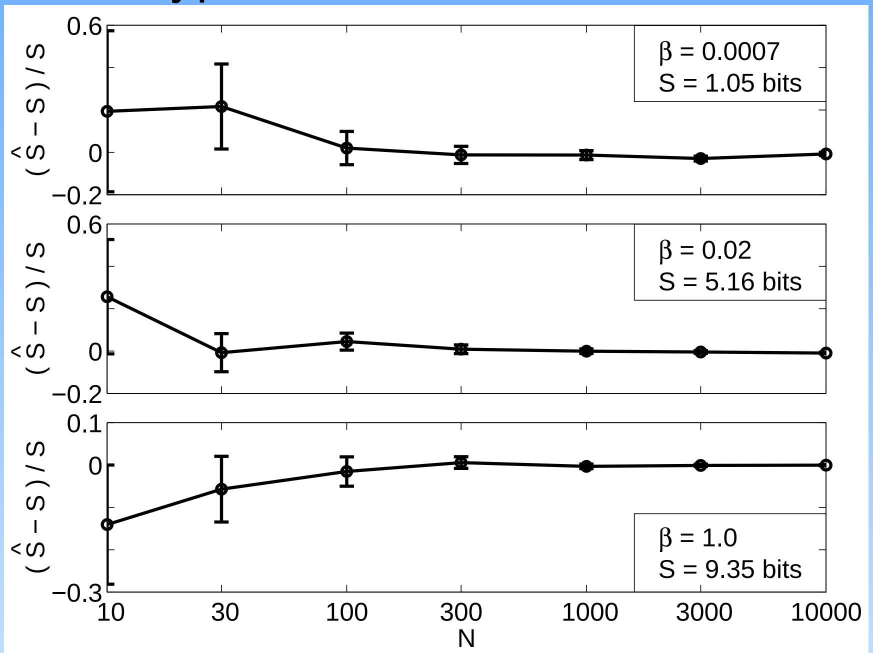
$$\widehat{S^m} = \frac{\int d\xi\, \rho(\xi, \{n_i\}) \langle\, S^m[n_i]\,\rangle_{\beta(\xi)}}{\int d\xi\, \rho(\xi, [n_i])}$$

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

- Smaller β means larger allowed volume in the space of $\{q_i\}$. Thus averaging over β is *Bayesian model selection*.
- $\langle \delta^2 S\rangle$ is dominated by $\langle \delta^2 \xi\rangle$, which is small if a particular β (model) dominates (is "selected")

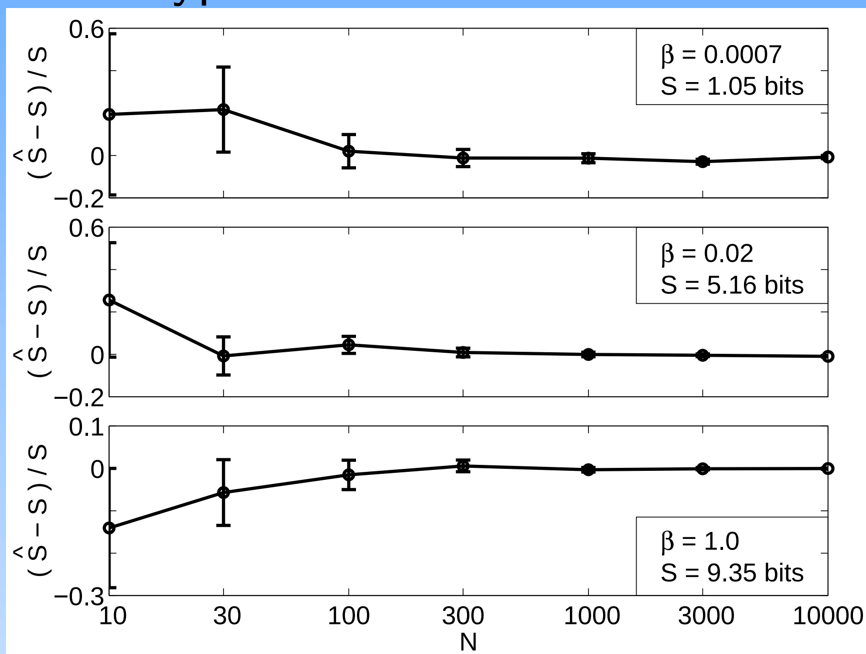
First attempts to estimate entropy

Typical distributions

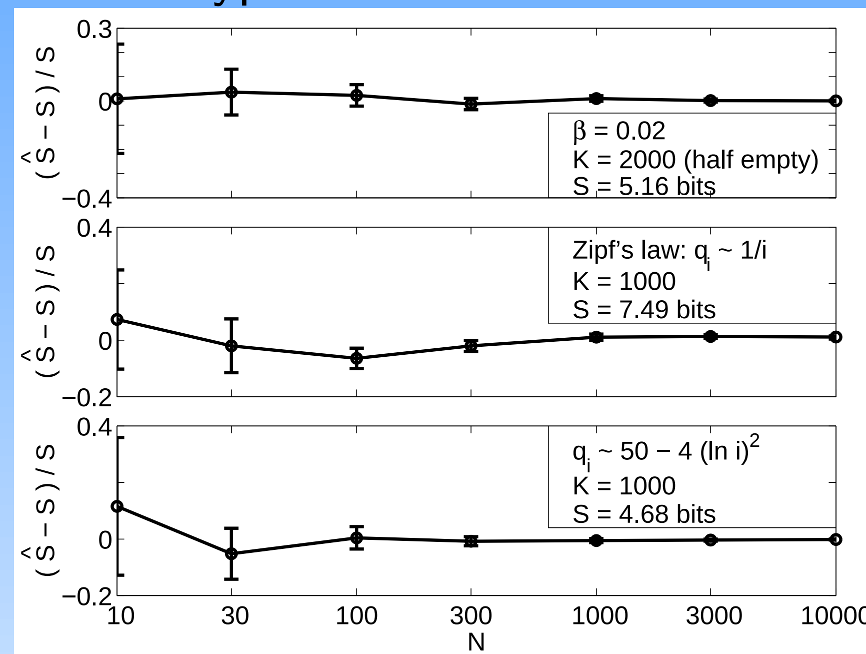


First attempts to estimate entropy

Typical distributions

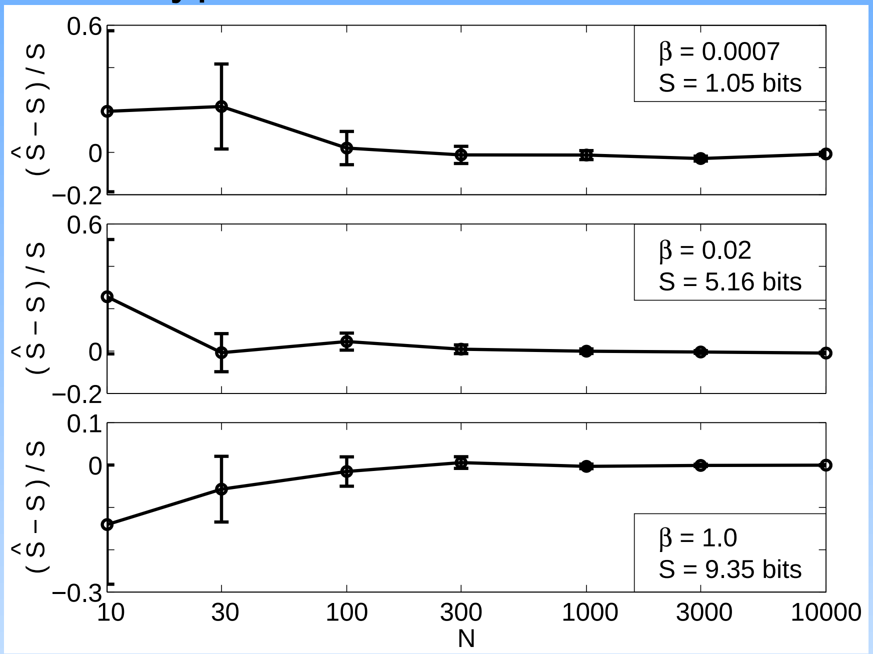


Atypical distributions

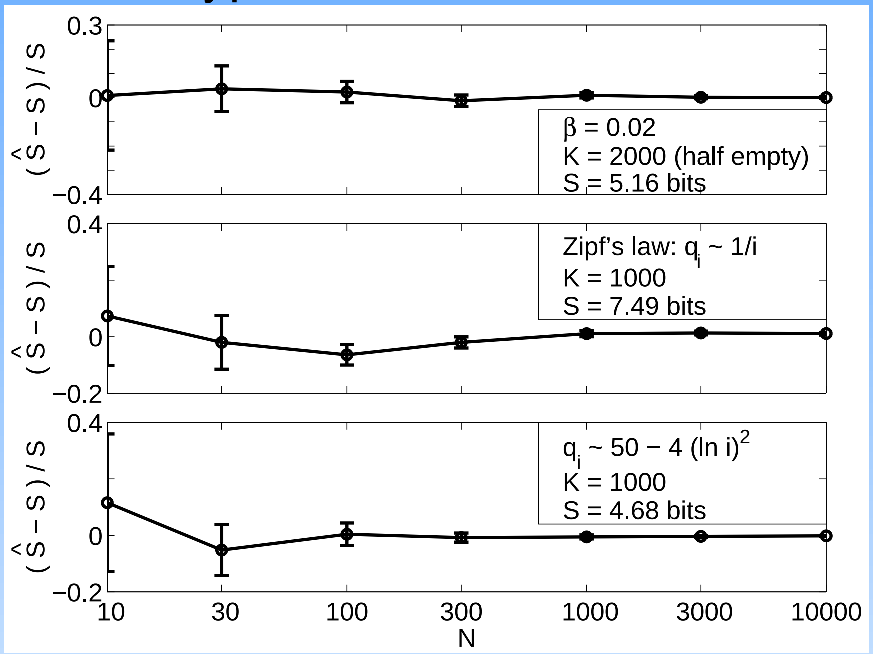


First attempts to estimate entropy

Typical distributions



Atypical distributions



Supports understanding that smoothness = speed of decay of Zipf plot.

Estimating entropy: first observations

- Relative error $\sim 10\%$ at N as low as 30 for $K = 1000$.
- Reliable estimation of error (posterior variance).
- *Little bias*, as it should be. Exception: too smooth distributions.
- Key point: *learn entropies directly without finding* $\{q_i\}$*!*
- The dominant β stabilizes for typical distributions; drifts down (to complex models) for rough ones and up (to simpler models) for too smooth cases.

Asymptotics

$$K \gg 1,\ \Delta \equiv N - K_{\mathrm{counts}>0} \gg 1$$

- saddle point works
- $\left.\frac{\partial^2(-\log\rho)}{\partial\xi^2}\right|_{\xi(\beta^*)} = \left[\frac{\partial^2(-\log\rho)}{\partial\beta^2}\frac{1}{(d\xi/d\beta)^2}\right]_{\beta^*} = \Delta + NO([\Delta/N]^2)$

Asymptotics

$K \gg 1,\ \Delta \equiv N - K_{\mathrm{counts}>0} \gg 1$

- saddle point works
- $\left.\frac{\partial^2(-\log\rho)}{\partial\xi^2}\right|_{\xi(\beta^*)} = \left[\frac{\partial^2(-\log\rho)}{\partial\beta^2}\frac{1}{(d\xi/d\beta)^2}\right]_{\beta^*} = \Delta + NO([\Delta/N]^2)$

$K, N \gg 1,\ \Delta \sim 1$

- $\widehat{S} \approx (C_\gamma - \ln 2) + 2\ln N - \psi_0(\Delta) + O(\frac{1}{N}, \frac{1}{K})$
- $(\widehat{\delta S})^2 \approx \psi_1(\Delta) + O(\frac{1}{N}, \frac{1}{K})$

Asymptotics

$K \gg 1$, $\Delta \equiv N - K_{\mathrm{counts}>0} \gg 1$

- saddle point works
- $\left.\frac{\partial^2(-\log\rho)}{\partial\xi^2}\right|_{\xi(\beta^*)} = \left[\frac{\partial^2(-\log\rho)}{\partial\beta^2}\frac{1}{(d\xi/d\beta)^2}\right]_{\beta^*} = \Delta + NO([\Delta/N]^2)$

$K, N \gg 1$, $\Delta \sim 1$

- $\widehat{S} \approx (C_\gamma - \ln 2) + 2\ln N - \psi_0(\Delta) + O(\frac{1}{N}, \frac{1}{K})$
- $\widehat{(\delta S)^2} \approx \psi_1(\Delta) + O(\frac{1}{N}, \frac{1}{K})$

Remember Ma's estimate!

Estimator: Properties

- K can be infinite
- Works for $\Delta \ll N$ if distribution is not atypically smooth.
- Δ matters, not K or N.
- The estimator is consistent.
- Thus correct if self-consistent for subsamples.
- When works, works for $N \sim 2^{S/2}$.

Estimator: Properties

- K can be infinite
- Works for $\Delta \ll N$ if distribution is not atypically smooth.
- Δ matters, not K or N.
- The estimator is consistent.
- Thus correct if self-consistent for subsamples.
- When works, works for $N \sim 2^{S/2}$.
- Selection of K by Bayesian integration not an option: small K means smaller phase space and *better* approximation.

Estimator: Synthetic test

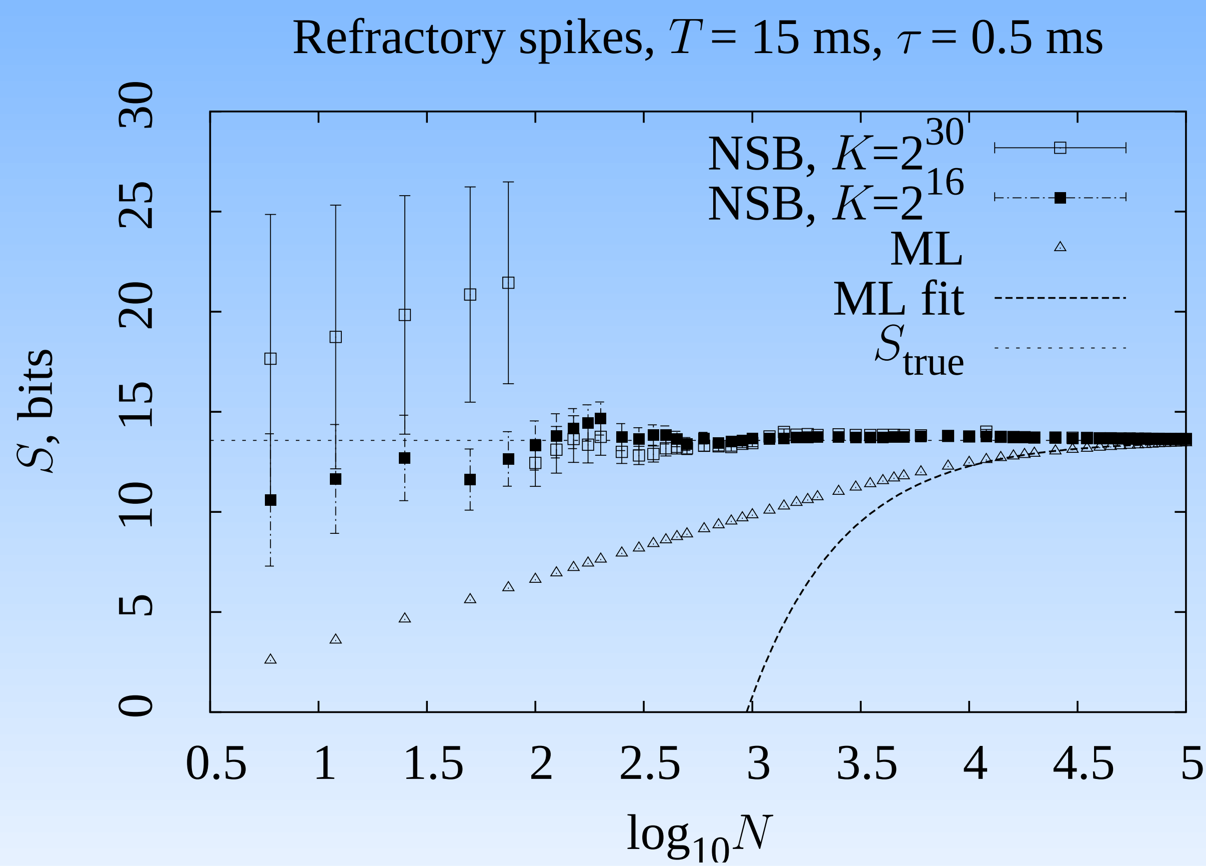
Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$.

Estimator: Synthetic test

Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$.
$K = 2^{30}$, $K_{\text{ref}} < 2^{16}$, $S = 13.57\text{bits}$.

Estimator: Synthetic test

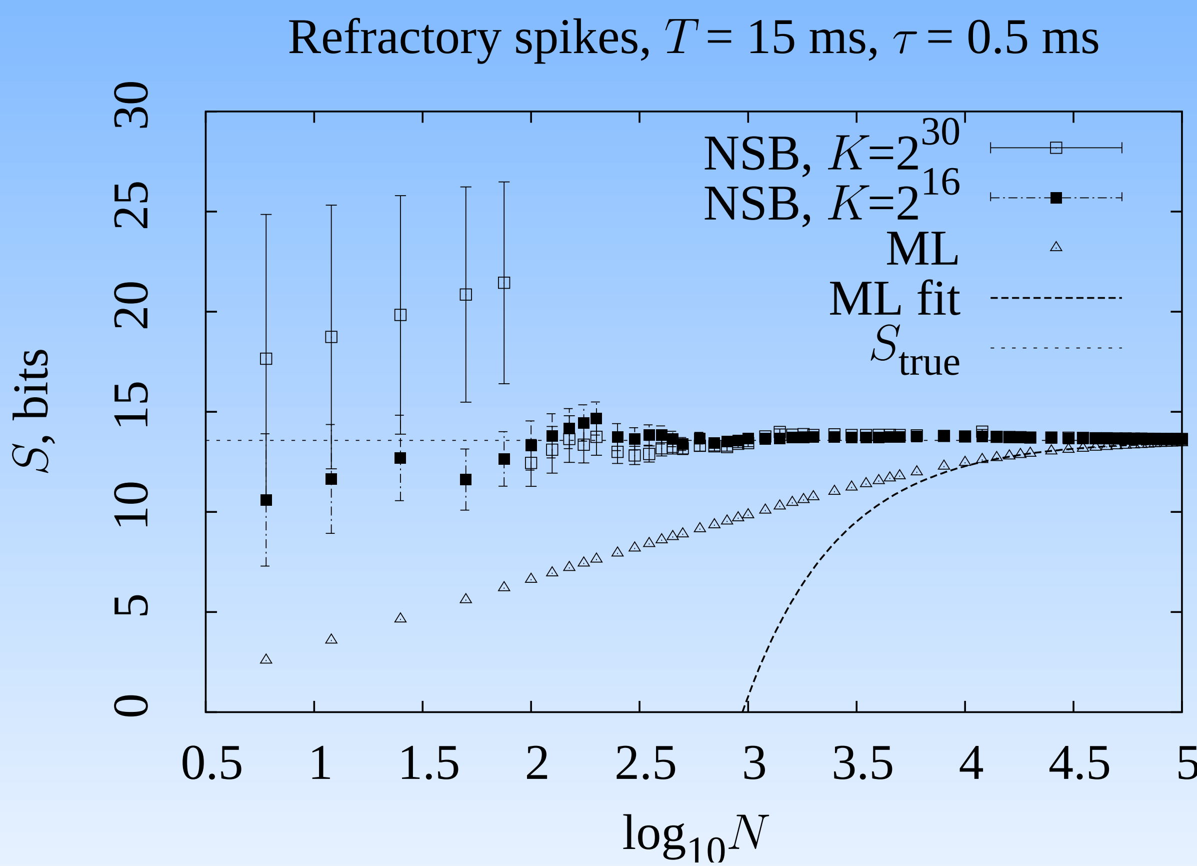
Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$. $K = 2^{30}$, $K_{\text{ref}} < 2^{16}$, $S = 13.57\text{bits}$.



True value reached within the error bars for $N^2 \sim 2^S$, when coincidences start to occur.

Estimator: Synthetic test

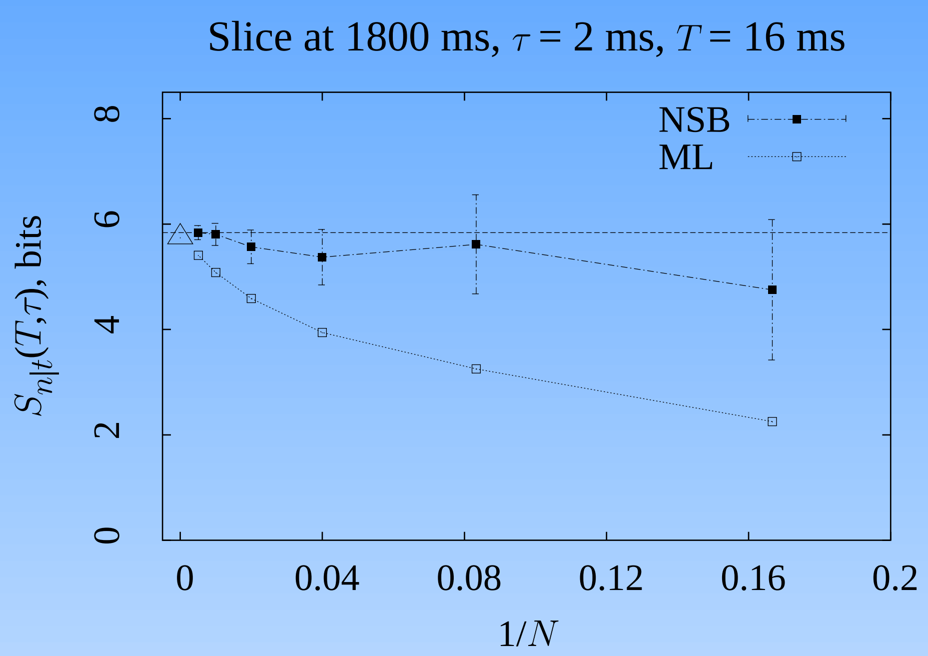
Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$. $K = 2^{30}$, $K_{\text{ref}} < 2^{16}$, $S = 13.57\text{bits}$.



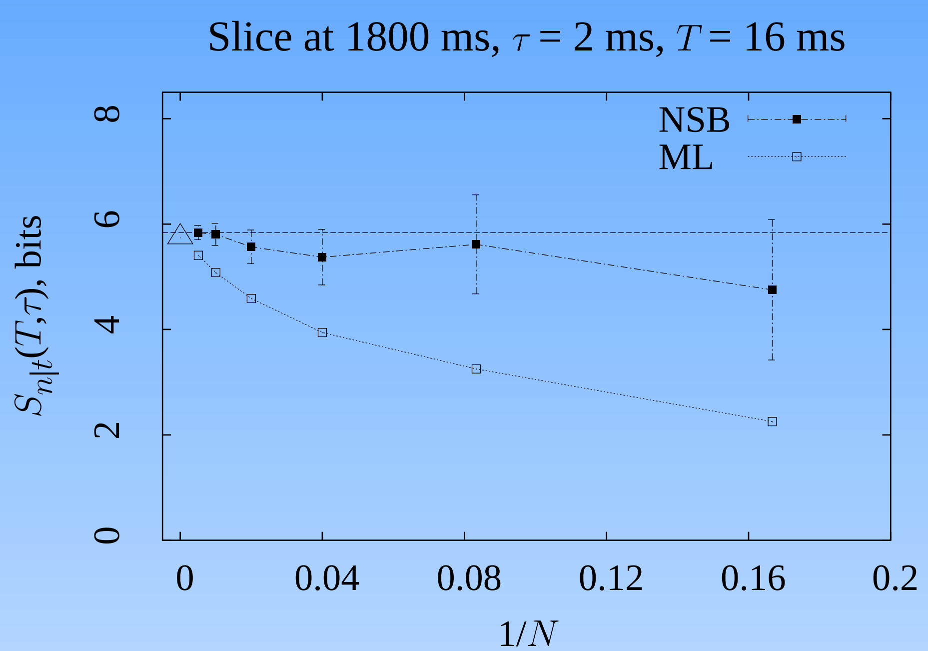
True value reached within the error bars for $N^2 \sim 2^S$, when coincidences start to occur.

Estimator is unbiased if it is consistent and agrees with itself for all N within error bars.

Natural data: Slice entropy vs. sample size

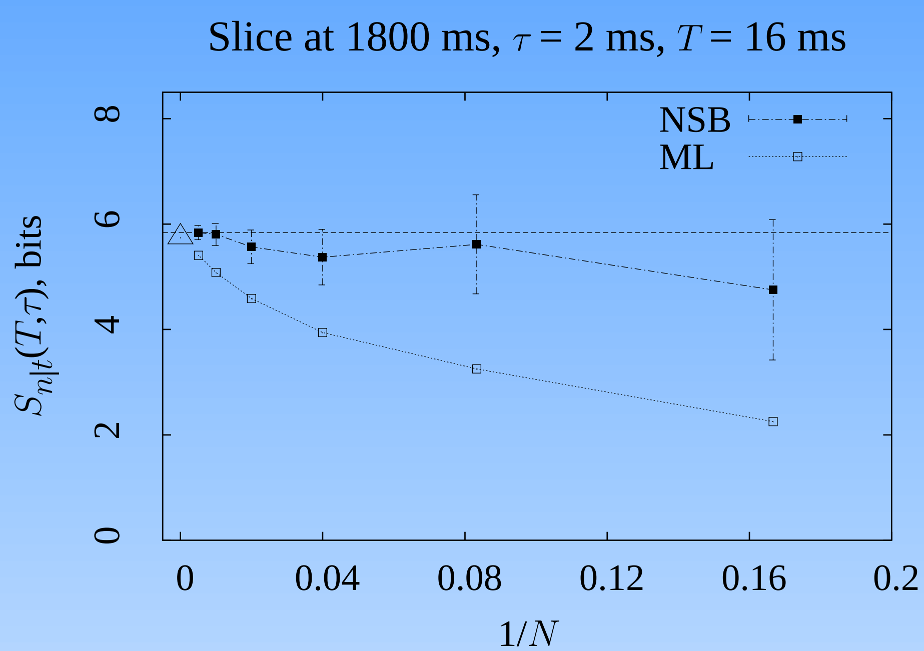


Natural data: Slice entropy vs. sample size



ML estimator converges with $\sim 1/N$ corrections.

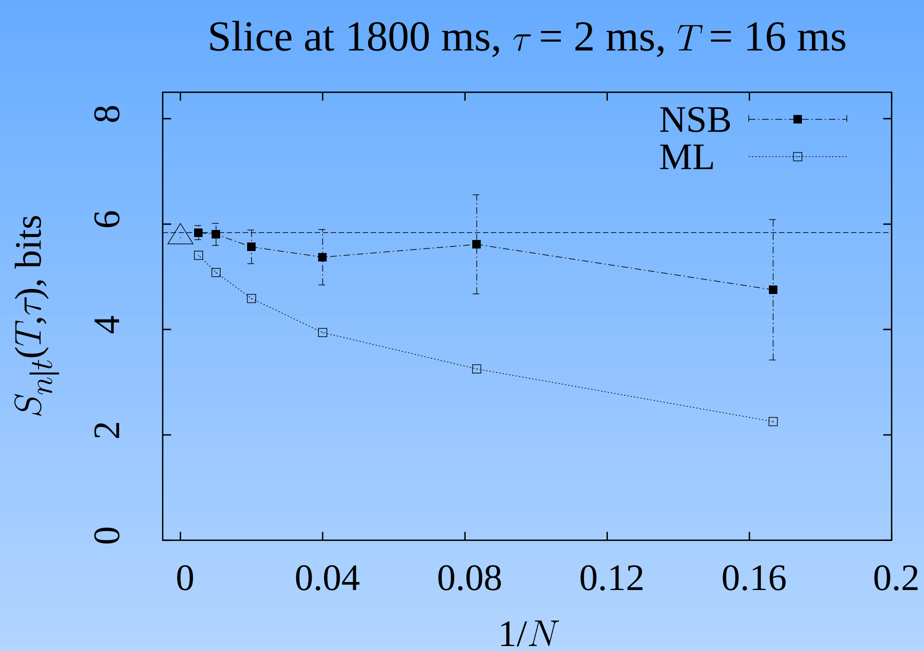
Natural data: Slice entropy vs. sample size

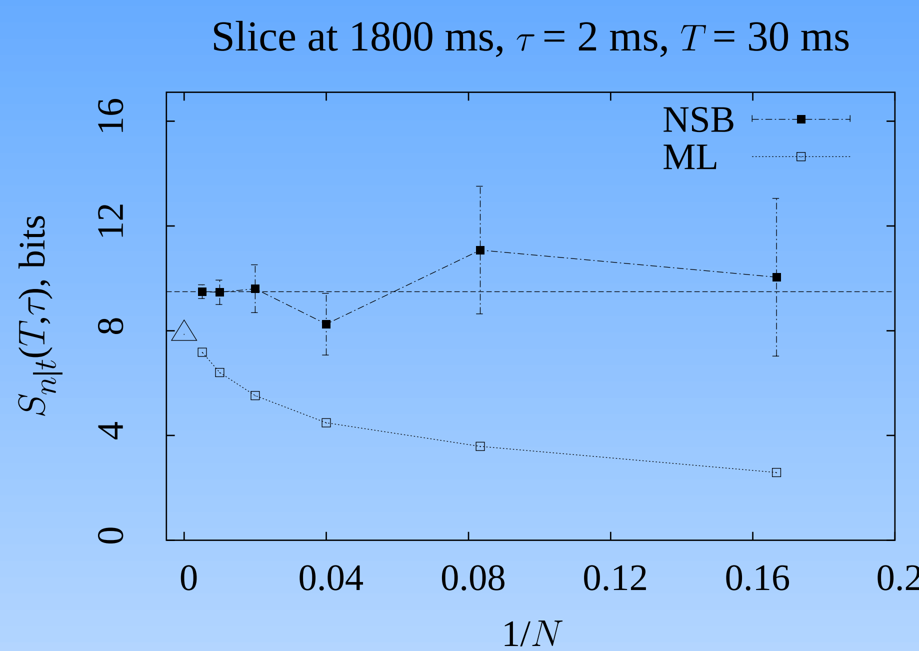


ML estimator converges with $\sim 1/N$ corrections.

NSB estimator is always within error bars.

Natural data: Slice entropy vs. sample size

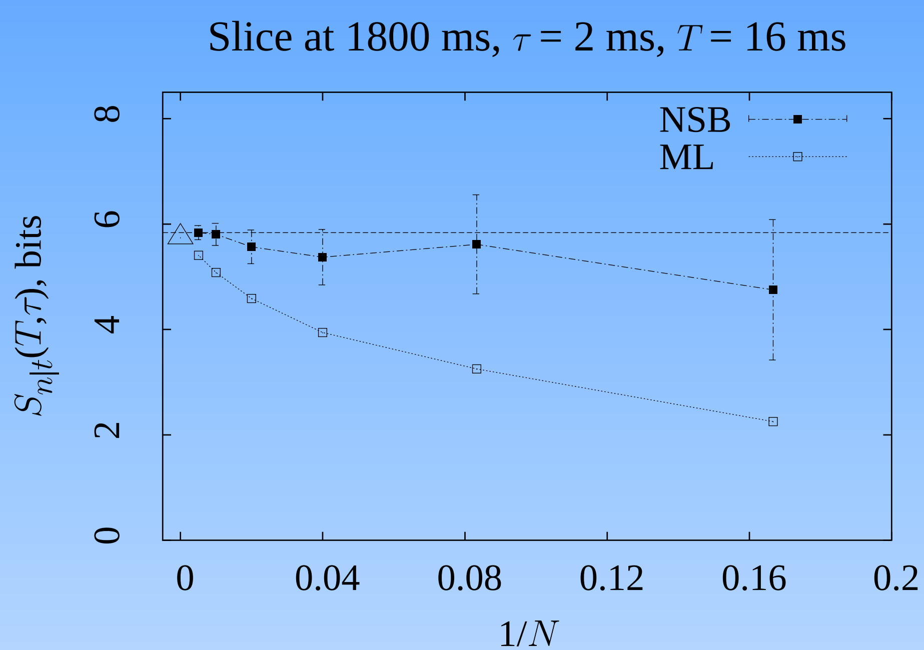


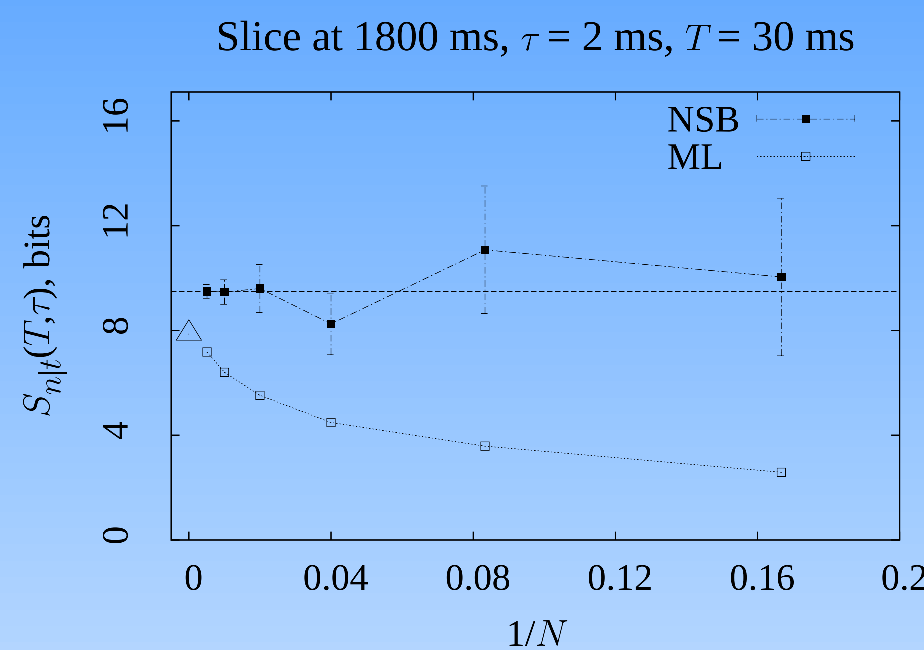


ML estimator converges with $\sim 1/N$ corrections.

NSB estimator is always within error bars.

Natural data: Slice entropy vs. sample size



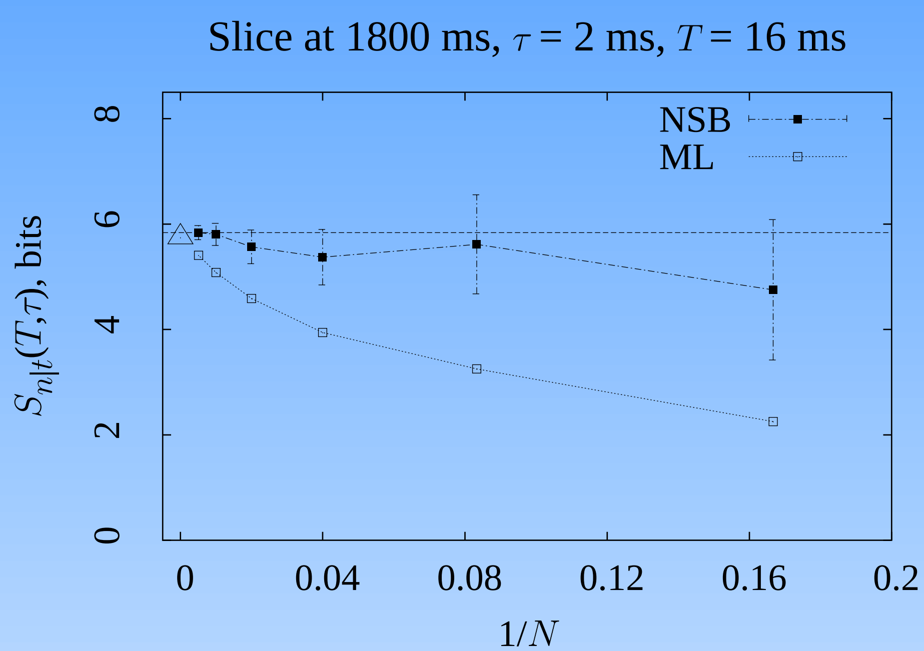


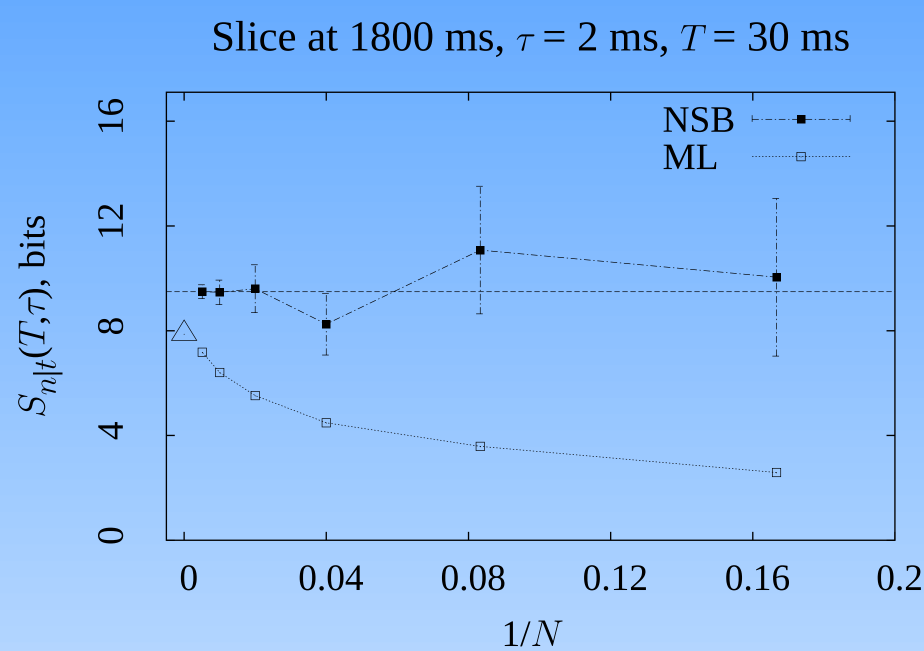
ML estimator converges with $\sim 1/N$ corrections.

NSB estimator is always within error bars.

ML estimator cannot be extrapolated.

Natural data: Slice entropy vs. sample size





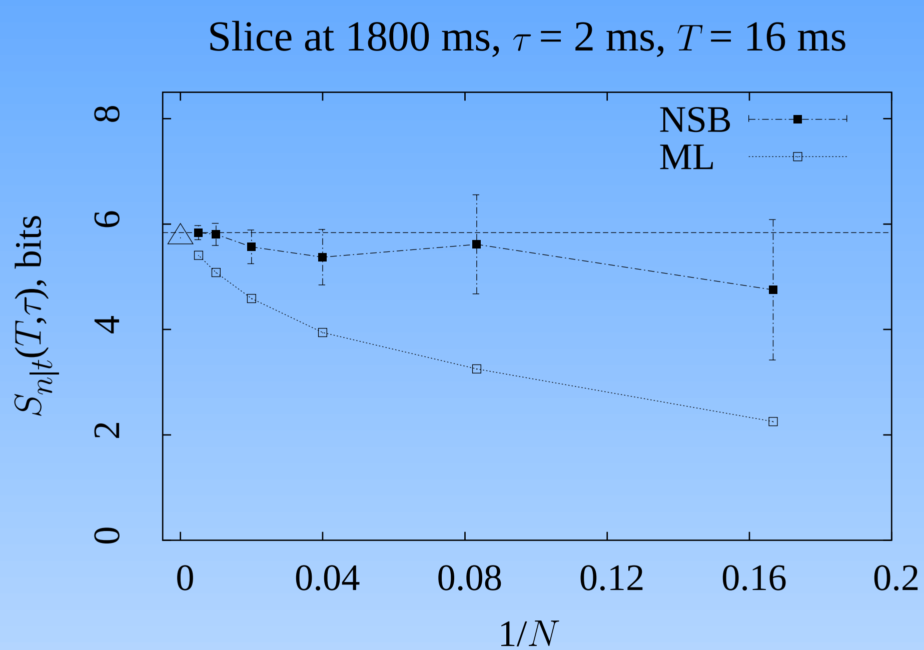
ML estimator converges with $\sim 1/N$ corrections.

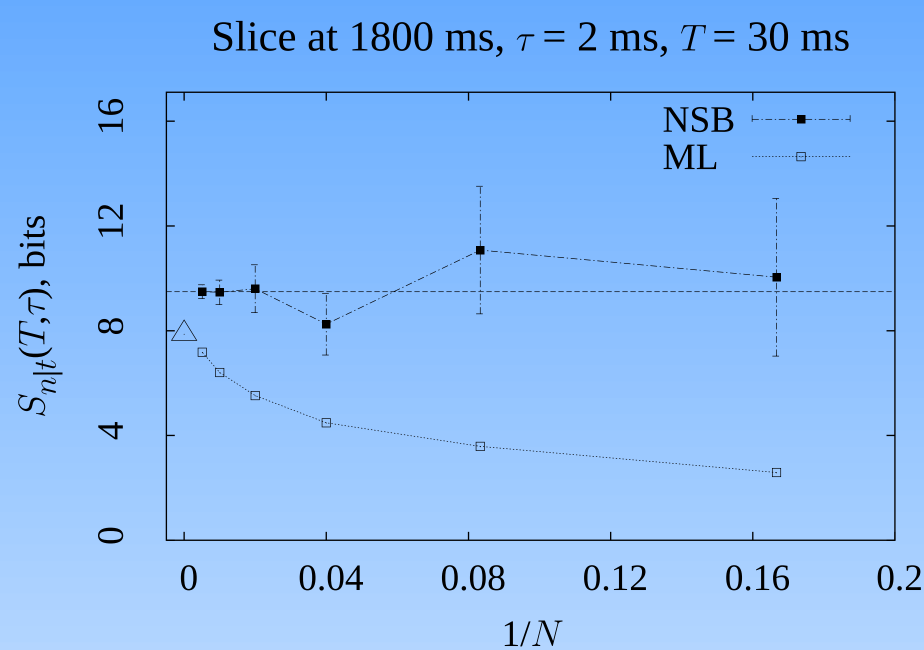
NSB estimator is always within error bars.

ML estimator cannot be extrapolated.

NSB estimator is always within error bars.

Natural data: Slice entropy vs. sample size





ML estimator converges with $\sim 1/N$ corrections.

NSB estimator is always within error bars.

ML estimator cannot be extrapolated.

NSB estimator is always within error bars.

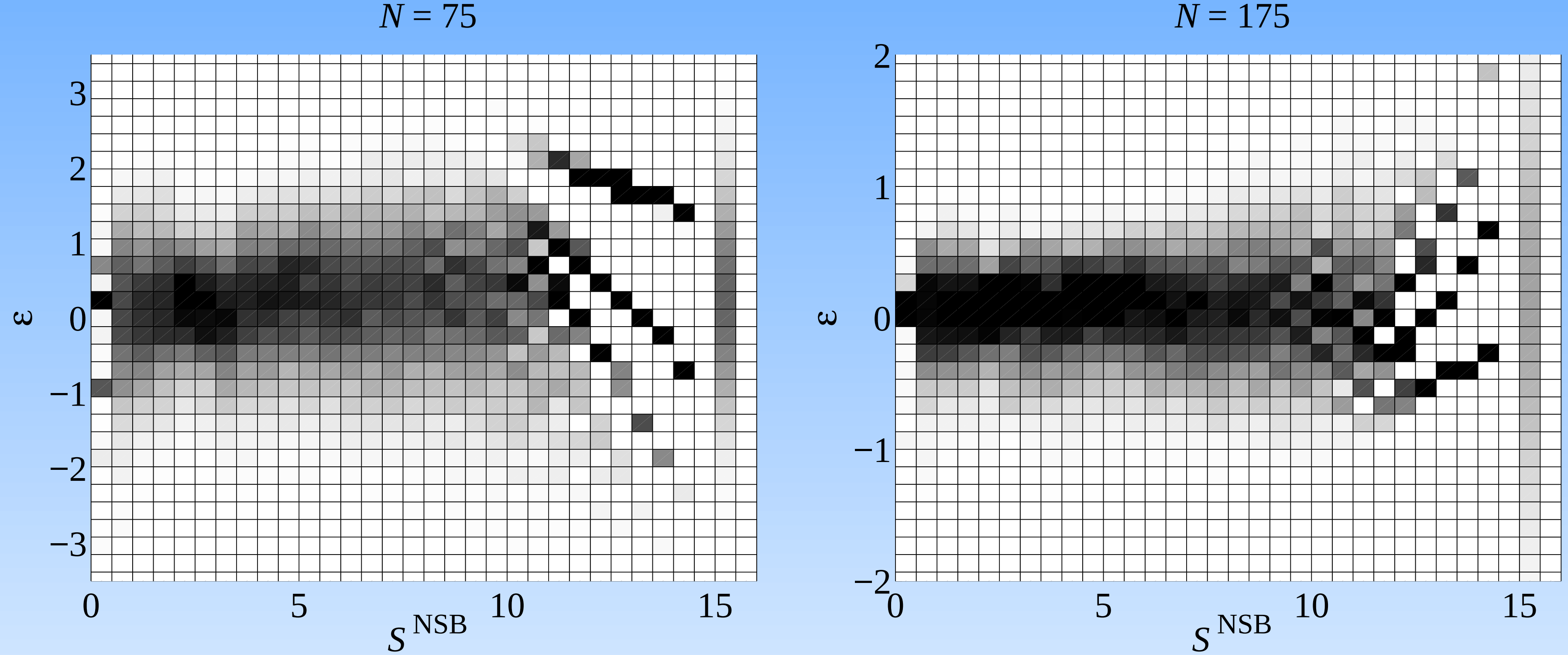
$(S^{\rm NSB} - S_{\rm ML})/\delta S^{\rm NSB}$ has zero mean if $S^{\rm ML}$ is reliably extrapolated ($N \gg 2^S$).

Natural data: Error vs. mean

$\epsilon(N) \equiv \frac{S^{\rm NSB}(N)-S}{\delta S^{\rm NSB}(N)} \approx \frac{S^{\rm NSB}(N)-S^{\rm NSB}(196)}{\delta S^{\rm NSB}(N)}$. Remember: $\log_2 196 \approx 7.5 \mathrm{bit}.$

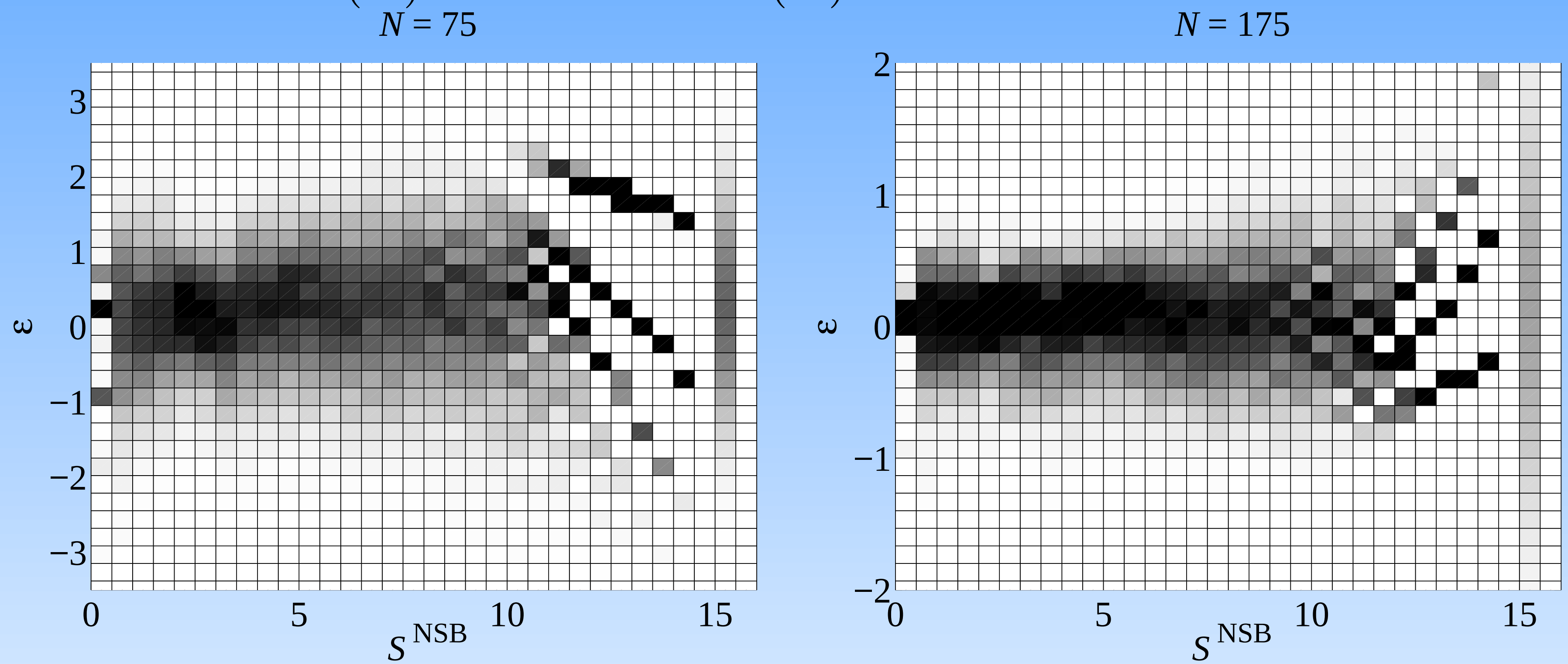
Natural data: Error vs. mean

$$\epsilon(N) \equiv \frac{S^{\rm NSB}(N) - S}{\delta S^{\rm NSB}(N)} \approx \frac{S^{\rm NSB}(N) - S^{\rm NSB}(196)}{\delta S^{\rm NSB}(N)}.$$ Remember: $\log_2 196 \approx 7.5 {\rm bit}.$



Natural data: Error vs. mean

$$\epsilon(N) \equiv \frac{S^{\rm NSB}(N) - S}{\delta S^{\rm NSB}(N)} \approx \frac{S^{\rm NSB}(N) - S^{\rm NSB}(196)}{\delta S^{\rm NSB}(N)}.$$ Remember: $\log_2 196 \approx 7.5 \mathrm{bit}$.



Almost no bias.

Empirical variance < 1 due to long tails in posterior, and $S \neq S^{\rm NSB}(196)$.

Bands are due to discrete nature of Δ.

Natural data: Hints of future results

Further work is needed to properly estimate error bars due to signal correlations.

Natural data: Hints of future results

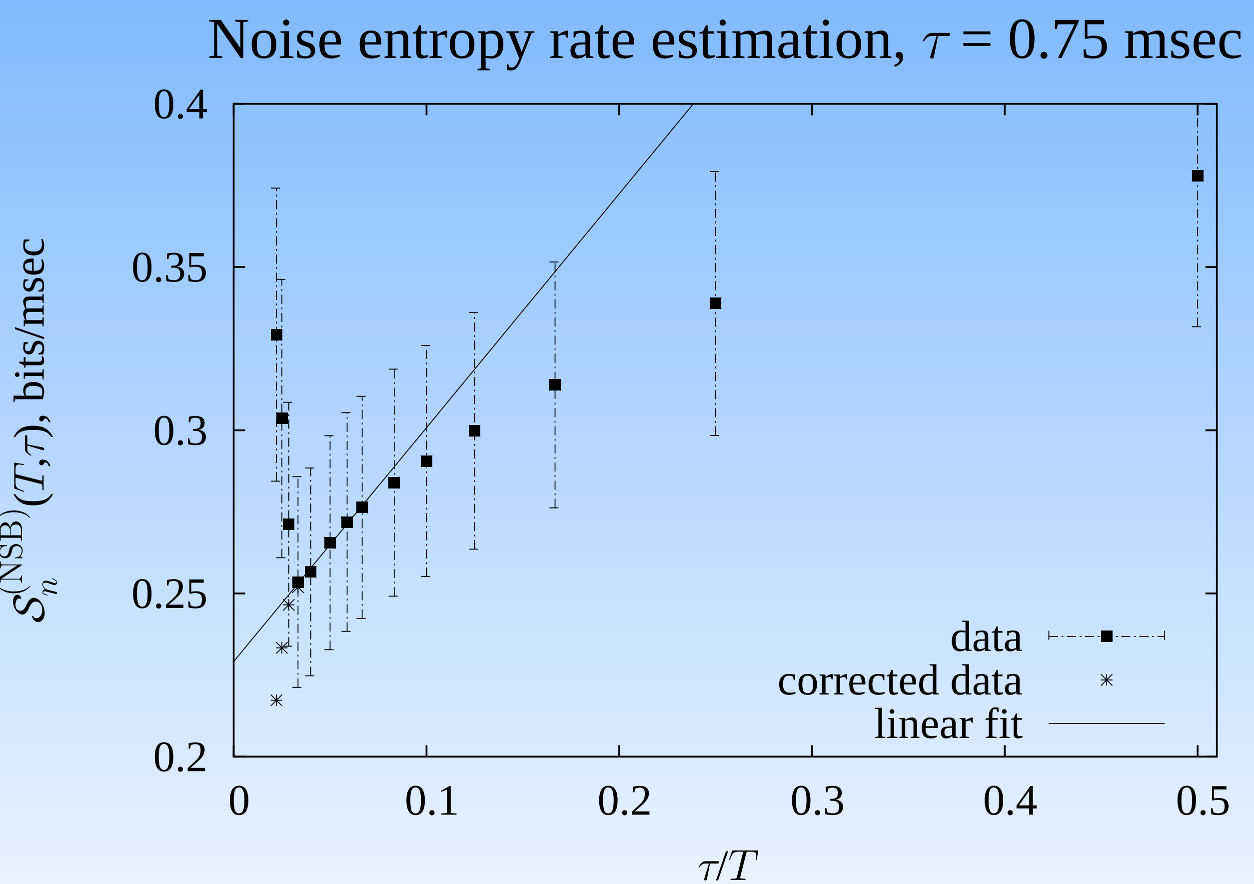
Further work is needed to properly estimate error bars due to signal correlations.

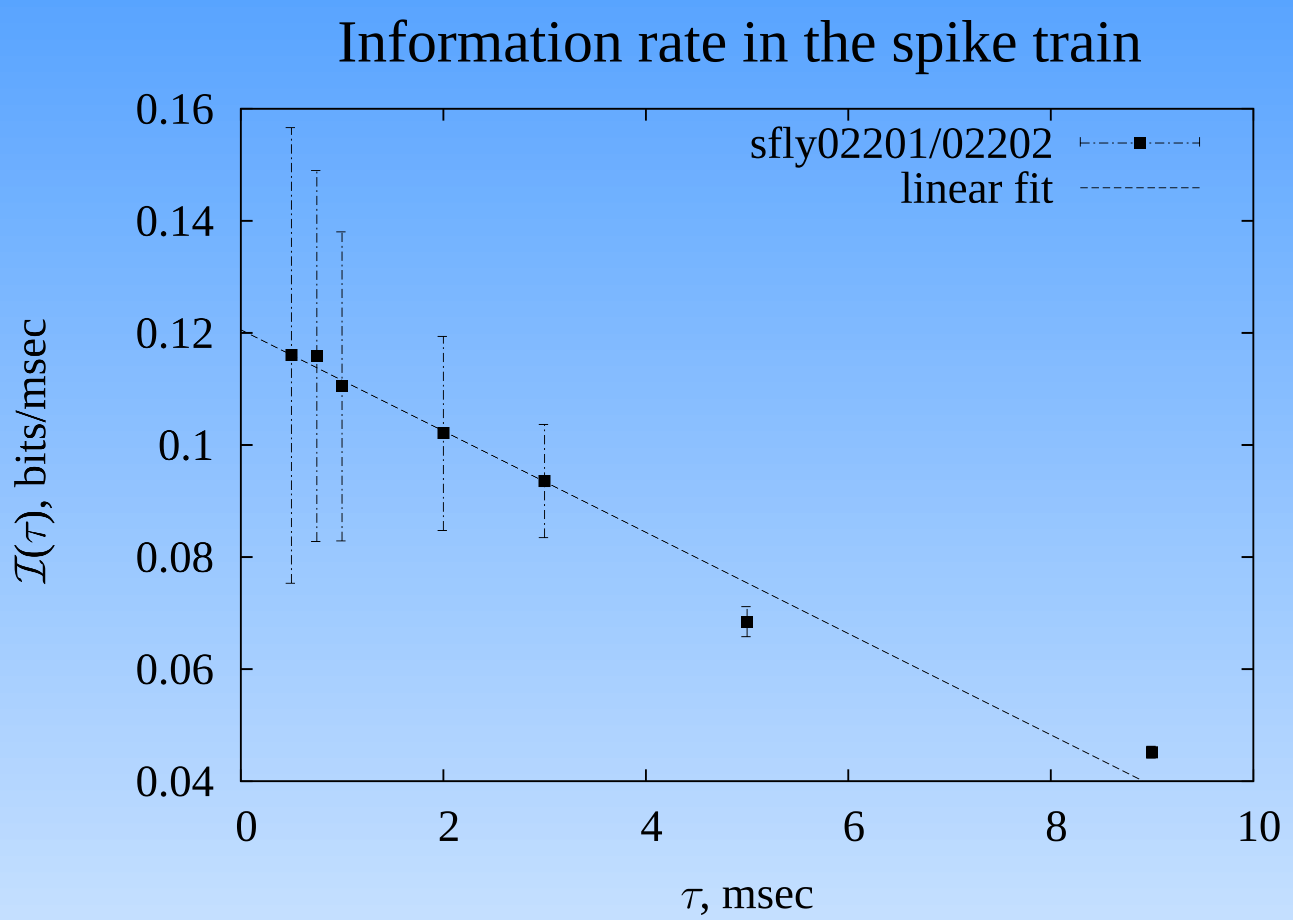
The fly in question is noisier than usual.

Natural data: Hints of future results

Further work is needed to properly estimate error bars due to signal correlations.

The fly in question is noisier than usual.

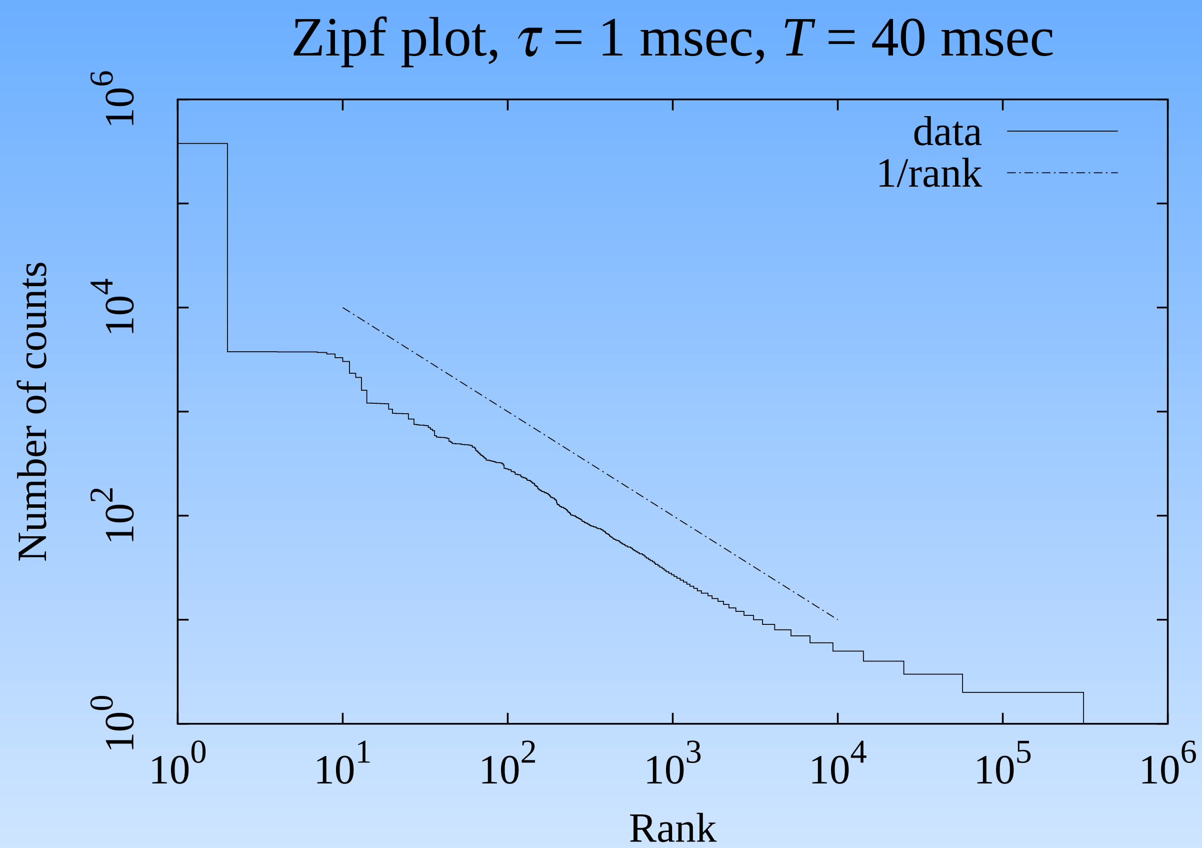



Information rate in the spike train
sfly02201/02202
linear fit
$\mathcal{I}(\tau)$, bits/msec
τ, msec
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0
2
4
6
8
10

Conclusions

- Found new entropy estimator.
- Works in Ma regime.
- Produces error bars.
- Know if we should trust it.
- Neural data seems to be well matched to the estimator

For amusement



Do not underestimate difficulty of working on real data!