Entropy and information of undersampled probability distributions

Ilya Nemenman
William Bialek, 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 Estimation information contents of spike trains, genomic sequences.

Talk outline

Problem setup Estimation information contents of spike trains, genomic sequences.

Developing intuition Why is it so difficult to estimate entropies?

Talk outline

Problem setup Estimation information contents of spike trains, genomic sequences.

Developing intuition Why is it so difficult to estimate entropies?

The method An idea.

The method Analysis.

The method Asymptotics.

The method Synthetic experiments.

Talk outline

Problem setup Estimation information contents of spike trains, genomic sequences.

Developing intuition Why is it so difficult to estimate entropies?

The method An idea.

The method Analysis.

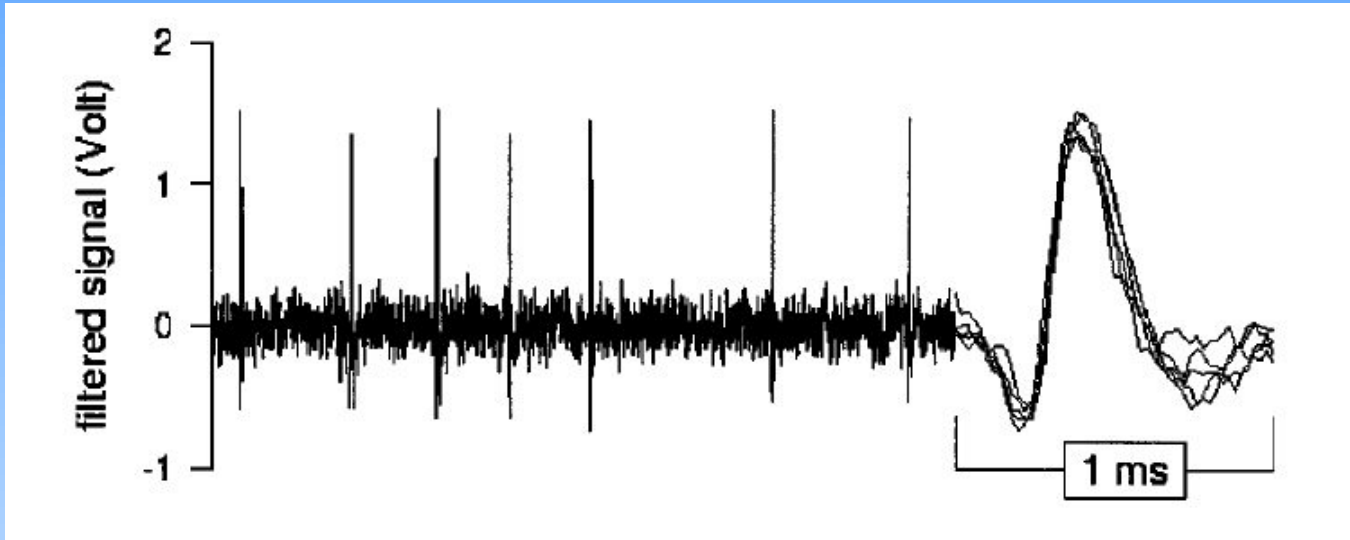
The method Asymptotics.

The method Synthetic experiments.

Applications Dealing with undersampling in neural data.

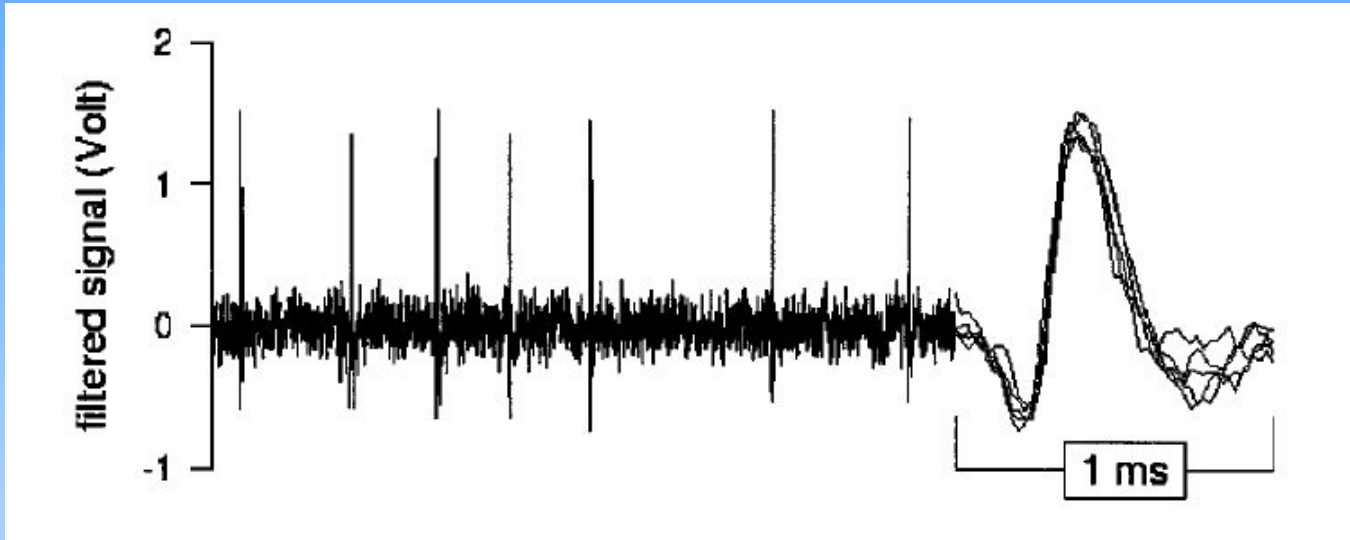
Applications Hints at future results.

Neurophysiological recordings



Strong et al., 1998

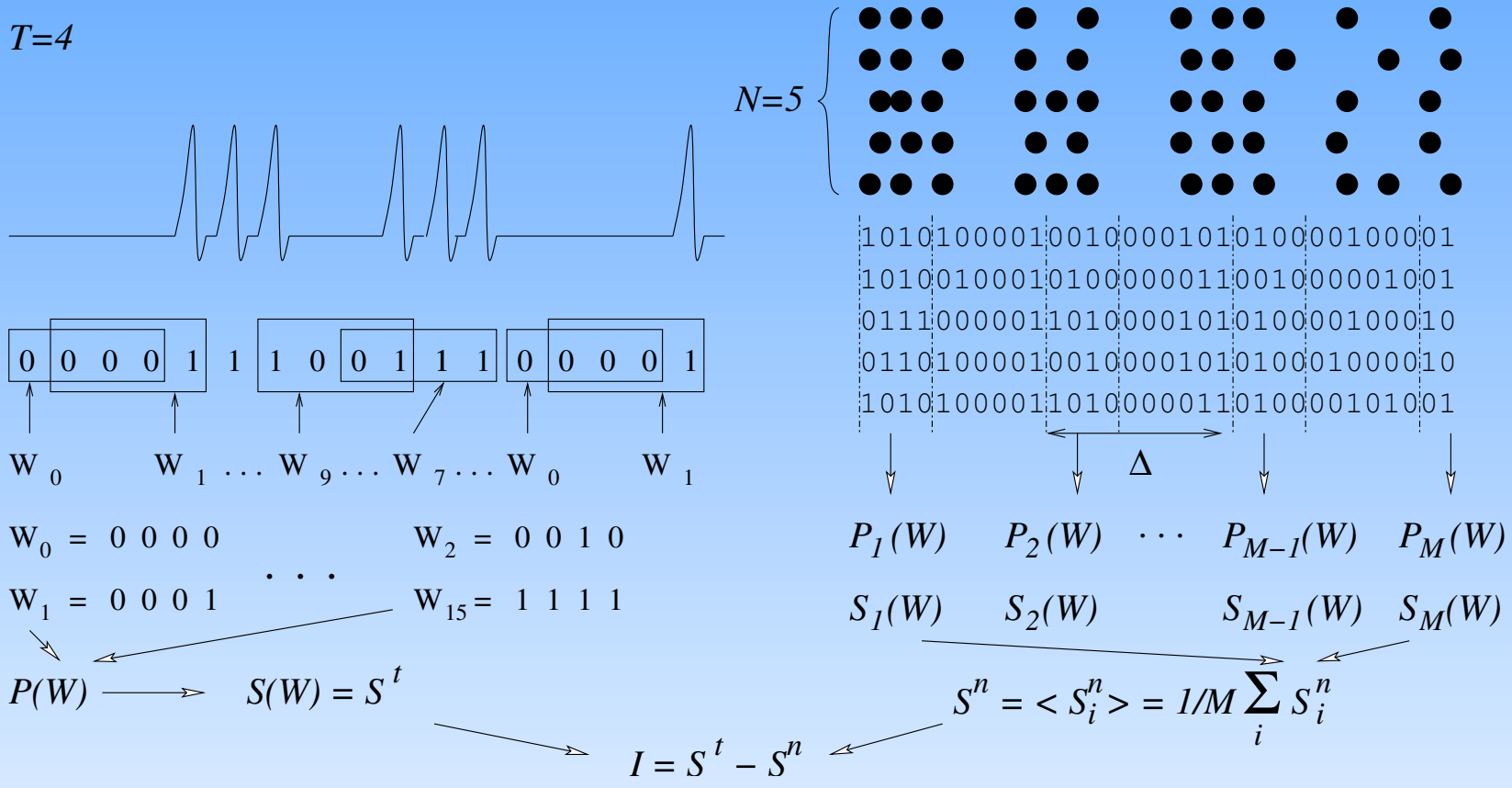
Neurophysiological recordings



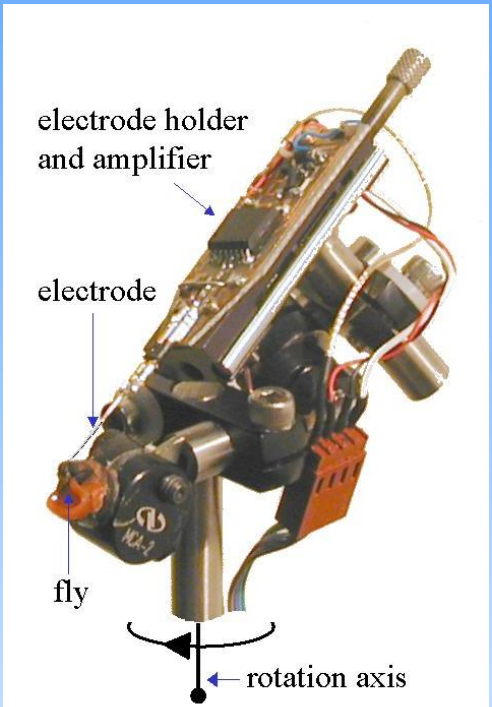
Strong et al., 1998

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

Estimating information rate in spike trains



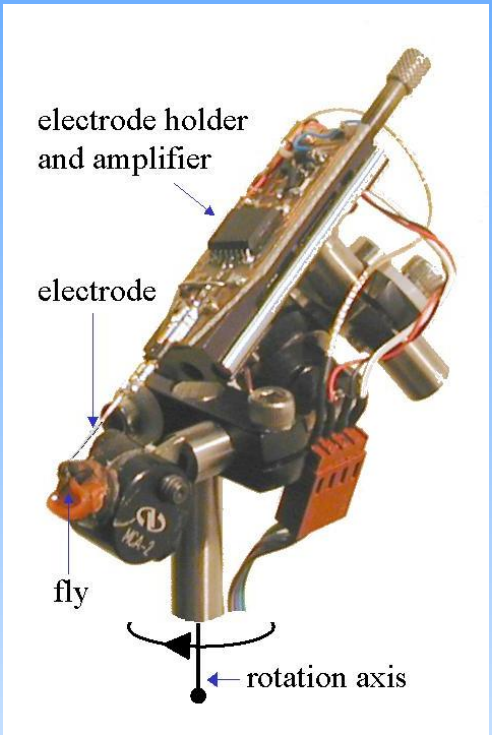
Experimental setup



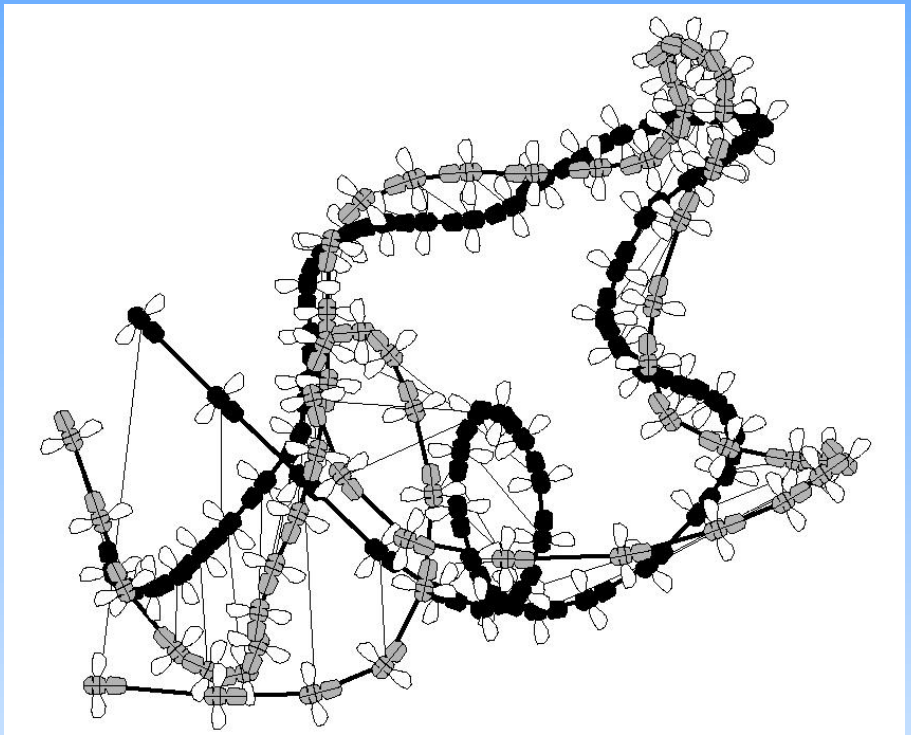
Lewen, Bialek, and de Ruyter

van Steveninck, 2001

Experimental setup



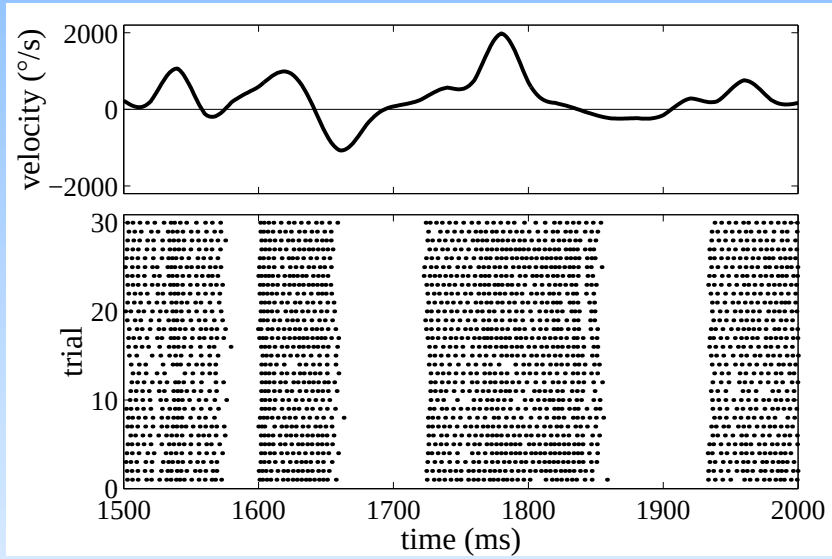
Lewen, Bialek, and de Ruyter van Steveninck, 2001



Bialek and de Ruyter van Steveninck, 2002, Land and Collett 1974

Recordings and problems

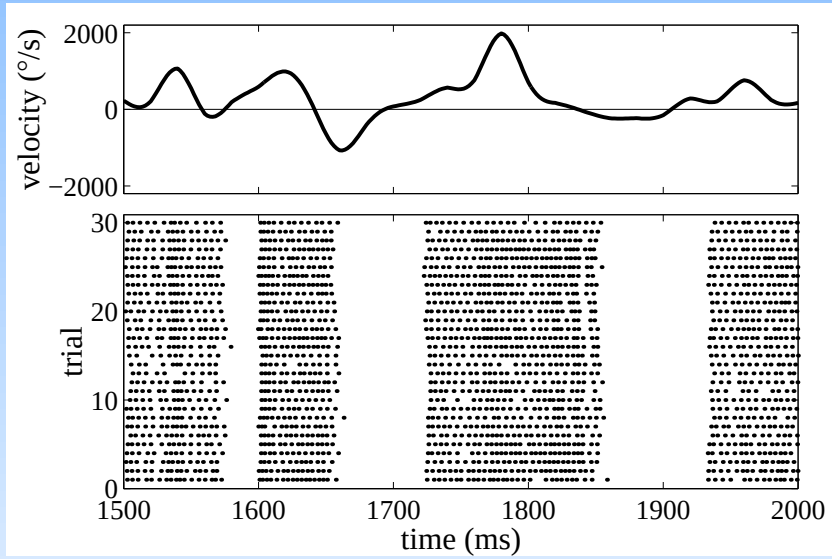
100–200 repeats of 5–10 s roller coasters rides



Recordings and problems

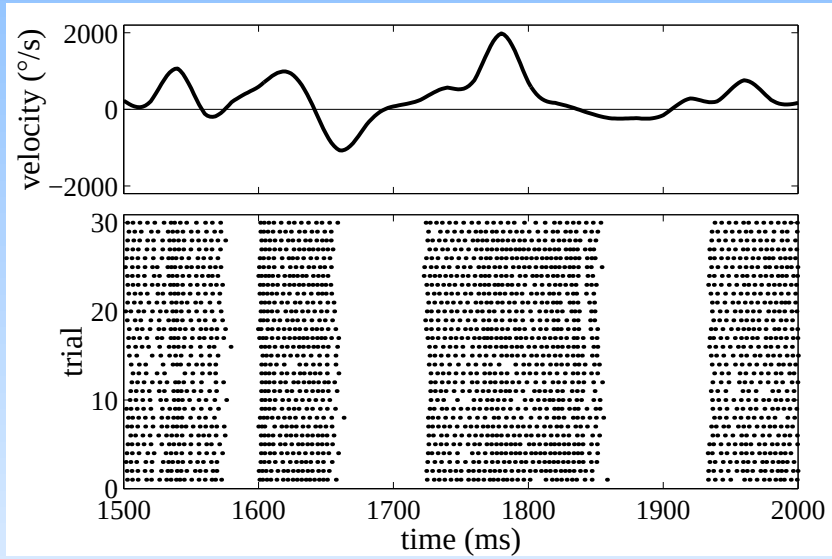
100–200 repeats of 5–10 s roller coasters rides

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



Recordings and problems

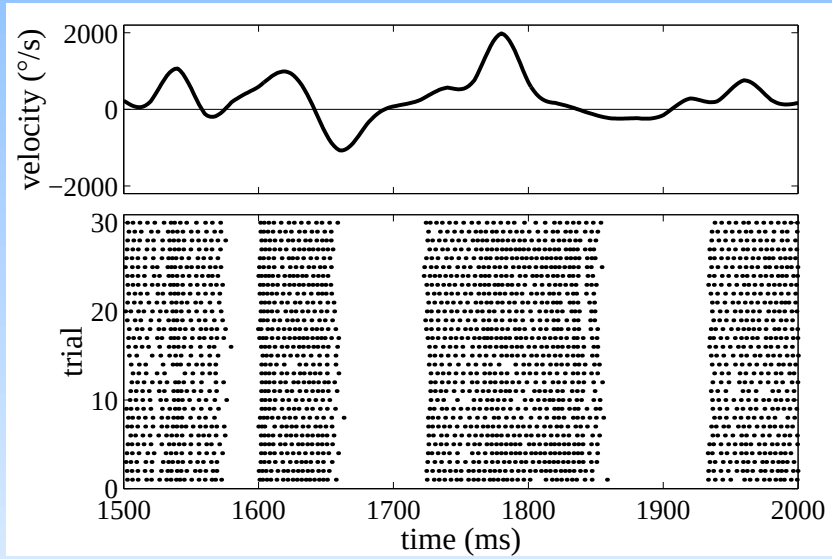
100–200 repeats of 5–10 s roller coasters rides



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

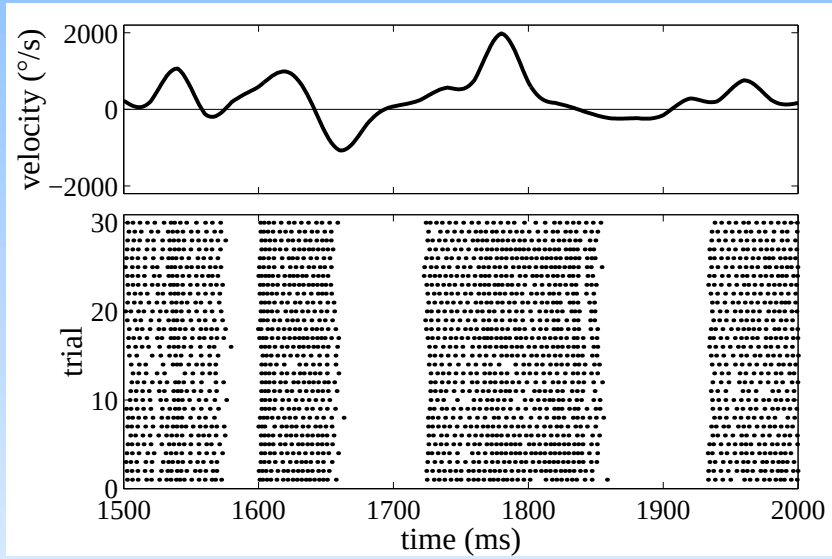
100–200 repeats of 5–10 s roller coasters rides



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

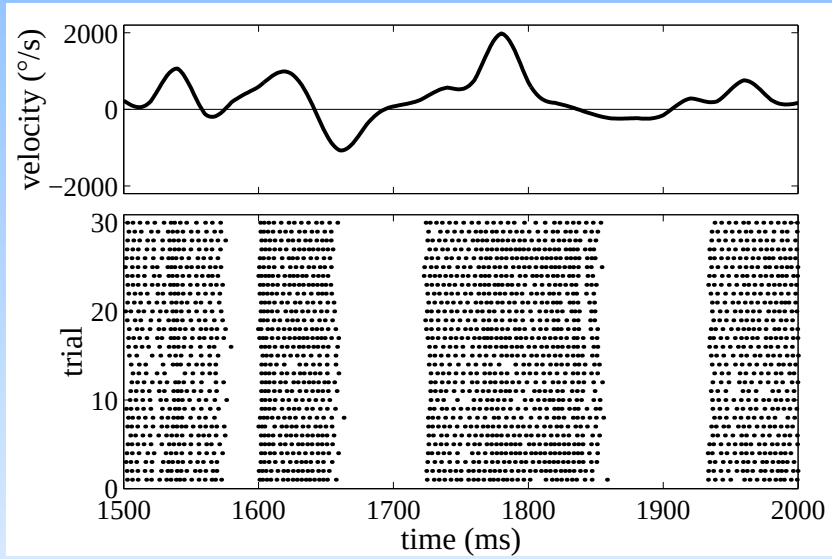
100–200 repeats of 5–10 s roller coasters rides



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.

Genomics analysis

$$GCCTA\overbrace{ACCGT}^{N}\underbrace{GGTCCA}_{D}\overbrace{TATATA}^{M}AGGAA$$

Genomics analysis

$$GCCTA\overbrace{ACCGT}^{N}\underbrace{GGTCCA}_{D}\overbrace{TATATA}^{M}AGGAA$$

Estimate mutual information $I(M, N; D)$.

Genomics analysis

$$GCCTA\overbrace{ACCGT}^{N}\underbrace{GGTCCA}_{D}\overbrace{TATATA}^{M}AGGAA$$

Estimate mutual information $I(M, N; D)$.

Study predictability properties.

Genomics analysis

$$GCCTA\overbrace{ACCGT}^{N}\underbrace{GGTCCA}_{D}\overbrace{TATATA}^{M}AGGAA$$

Estimate mutual information $I(M, N; D)$.

Study predictability properties.

Search for motifs.

Genomics analysis

$$GCCTA\overbrace{ACCGT}^{N}\underbrace{GGTCCA}_{D}\overbrace{TATATA}^{M}AGGAA$$

Estimate mutual information $I(M, N; D)$.

Study predictability properties.

Search for motifs.

Run IB and extract predictive features.

Why is it difficult to estimate entropies?

Suppose ϵ of the probability mass is in K (unknown) number of bins.

Why is it difficult to estimate entropies?

Suppose ϵ of the probability mass is in K (unknown) number of bins. This may contribute $\delta S = \epsilon \log_2 K$ to entropy.

Why is it difficult to estimate entropies?

Suppose ϵ of the probability mass is in K (unknown) number of bins. This may contribute $\delta S = \epsilon \log_2 K$ to entropy. $\forall \epsilon \ll 1, M \gg 1, \quad \exists K : \delta S > M.$

Why is it difficult to estimate entropies?

Suppose ϵ of the probability mass is in K (unknown) number of bins. This may contribute $\delta S = \epsilon \log_2 K$ to entropy. $\forall \epsilon \ll 1, M \gg 1, \quad \exists K : \delta S > M.$

$$\{Q_1, Q_2\} \longrightarrow \{n_1, n_2\}$$

$$\longrightarrow \{Q_1 + \delta, Q_2 - \delta\} \longrightarrow S - S_{\text{true}} < 0$$

Last step due to nonlinearity of $\log_2 P$.

Why is it difficult to estimate entropies?

Suppose ϵ of the probability mass is in K (unknown) number of bins. This may contribute $\delta S = \epsilon \log_2 K$ to entropy. $\forall \epsilon \ll 1, M \gg 1, \quad \exists K : \delta S > M.$

$$\{Q_1, Q_2\} \longrightarrow \{n_1, n_2\}$$

$$\longrightarrow \{Q_1 + \delta, Q_2 - \delta\} \longrightarrow S - S_{\text{true}} < 0$$

Last step due to nonlinearity of $\log_2 P$.

More in Paninski, 2003, Grassberger 2003.

There is no unbiased finite variance estimator of entropy.

Undersampling: metric cases

(weather, stocks,. . .)

Possible outcomes	$x, a \leq x \leq b$
Probability density	$Q(x)$
Observed data	$x_\mu,\ \mu = 1 \ldots N$
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

Undersampling: non–metric cases

(languages, bioinformatics,. . .)

Discrete outcomes (bins)	$i,\ i = 1 \ldots K$
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 $N \ll K$

We choose . . .

(for discrete case)

We choose . . .

(for discrete case)

1. Define smoothness as high entropy or low mutual information distributions.

We choose . . .

(for discrete case)

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

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

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

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

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

Some common choices:

Maximum likelihood $\beta \to 0$

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

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

Some common choices:

Maximum likelihood	$\beta \to 0$
Laplace's successor rule	$\beta = 1$

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

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

Some common choices:

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

Learning with nearly uniform priors

(ultra–local, Dirichlet priors)

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

Some common choices:

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

Numerics of the Dirichlet family

To generate distributions: Successively select each q_i according to

$$P(q_i) = B\left(\frac{q_i}{1-\sum_{j<i} q_j}; \beta, (K-i)\beta\right)$$

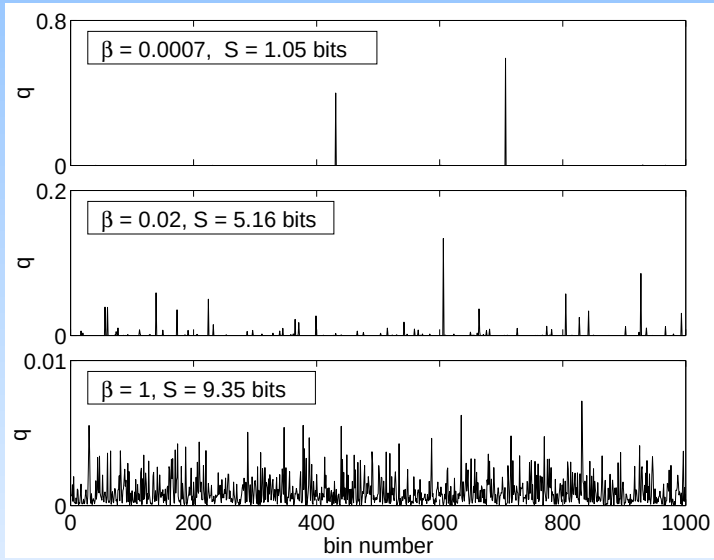
$$B(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

Numerics of the Dirichlet family

To generate distributions: Successively select each q_i according to

$$P(q_i) = B\left(\frac{q_i}{1-\sum_{j<i} q_j}; \beta, (K-i)\beta\right)$$

$$B(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$



Typical distributions ($K = 1000$). Note that the $\beta = 1$ distribution is very non–uniform, but has almost the maximum entropy (maybe reorder bins?)

Bayesian inference with Dirichlet priors

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

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

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

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

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

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

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

Equal pseudocounts added to each bin.

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

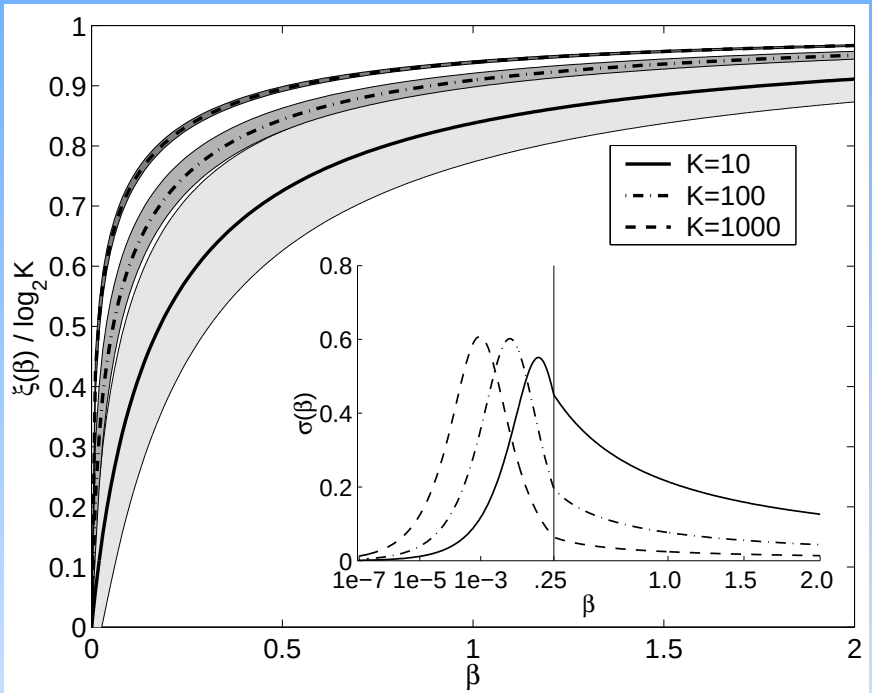
A problem: A priori entropy expectation

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

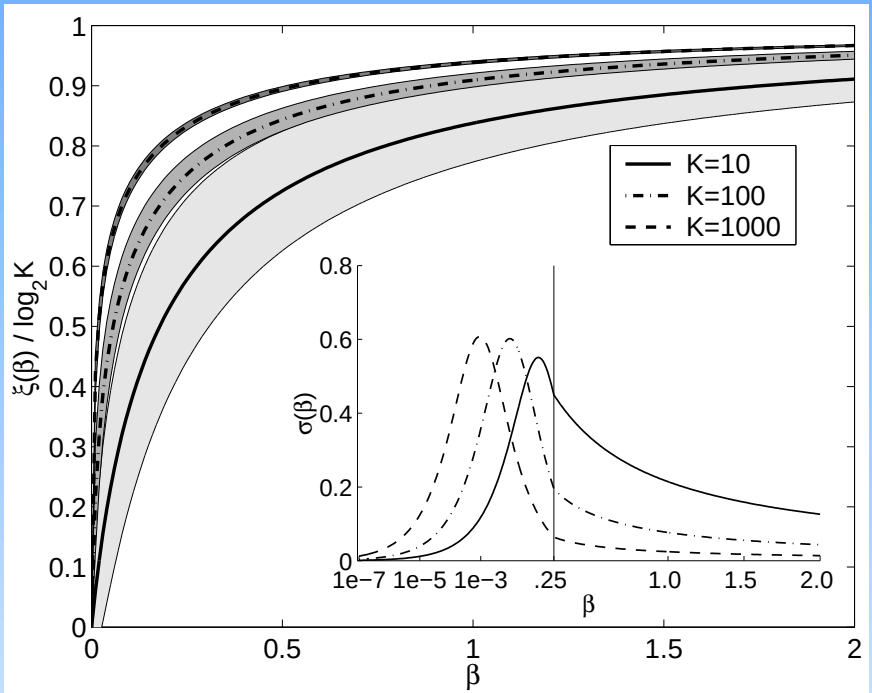
A problem: A priori entropy expectation

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

The problem: Analysis

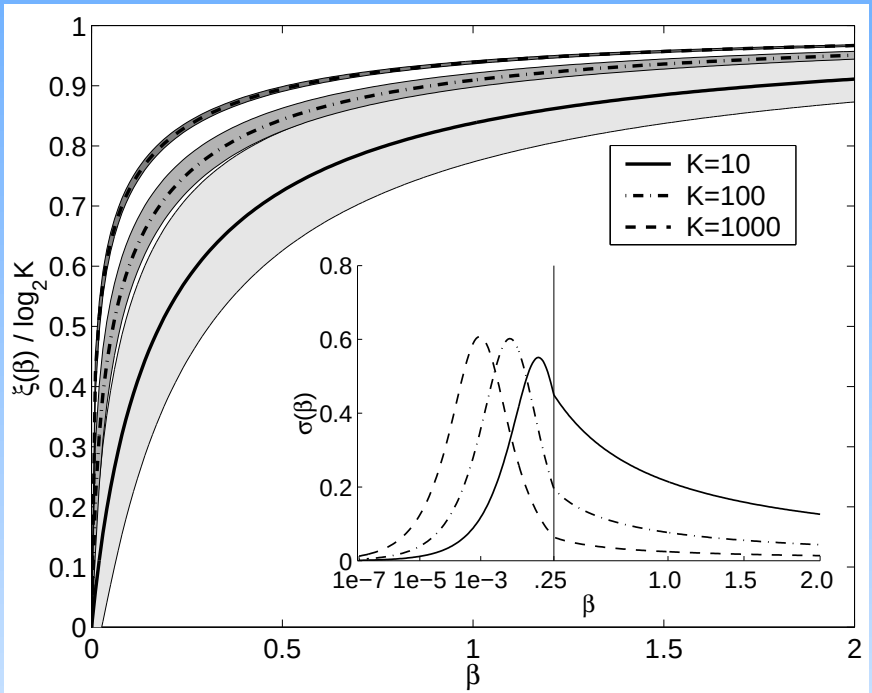


The problem: Analysis



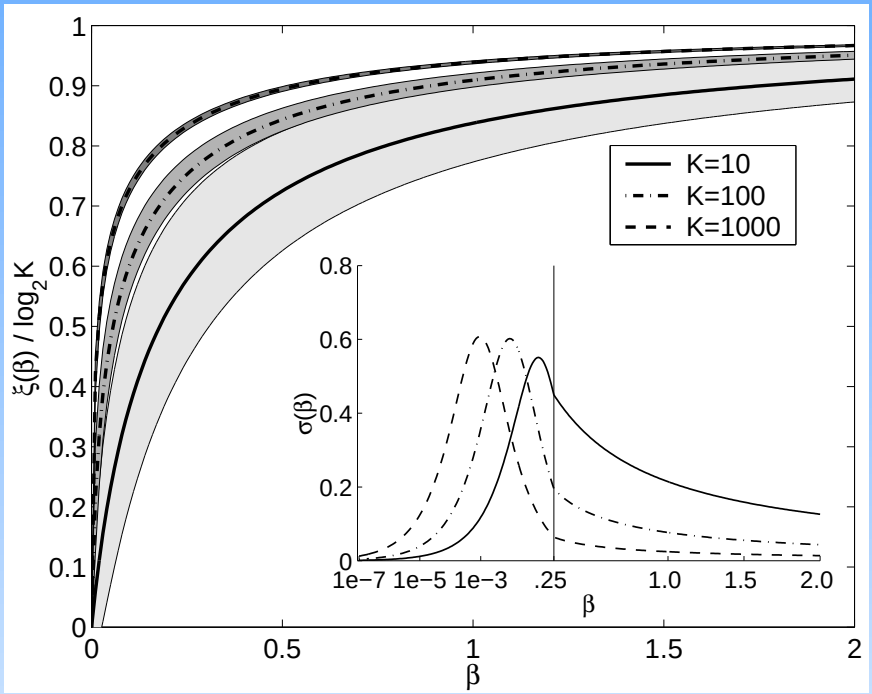
1. Because of the Jacobian of the $\{q_i\} \rightarrow S$ transformation, a priori distribution of entropy is strongly peaked.

The problem: Analysis



1. Because of the Jacobian of the $\{q_i\} \to S$ transformation, a priori distribution of entropy is strongly peaked.
2. *Narrow* peak: $\max \sigma(\beta) = 0.61\,\text{bits} \ll \log_2 K$ at $\beta \approx 1/K$; $\sigma(\beta) \propto 1/\sqrt{K\beta}$ for $K\beta \gg 1$; $\sigma(\beta) \propto \sqrt{K\beta}$ for $K\beta \ll 1$.

The problem: Analysis



1. Because of the Jacobian of the $\{q_i\} \to S$ transformation, a priori distribution of entropy is strongly peaked.
2. *Narrow* peak: $\max \sigma(\beta) = 0.61\,\text{bits} \ll \log_2 K$ 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, $\xi(\beta) = \log_2 K - O(K^0)$.

The problem

1. No a priori way to specify β.

The problem

1. No a priori way to specify β.
2. Choosing β fixes allowed "shapes" of $\{q_i\}$, and thus defines the a priori expectation of entropy.

The problem

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

The problem

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

Problems of common estimators

Maximum likelihood

Problems of common estimators

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

Maximum likelihood

Problems of common estimators

Maximum likelihood

$$\mathcal{P}_0(S) = \delta(S)$$
$$S = S_{\mathrm{ML}} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$$

Problems of common estimators

Maximum likelihood

$$\mathcal{P}_0(S) = \delta(S)$$
$$S = S_{\mathrm{ML}} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$$

(K^* is estimated ad hoc)

Problems of common estimators

Maximum likelihood

$$\mathcal{P}_0(S) = \delta(S)$$
$$S = S_{\mathrm{ML}} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$$
(K^* is estimated ad hoc)

Laplace and KT

Problems of common estimators

Maximum likelihood

$$\mathcal{P}_0(S) = \delta(S)$$
$$S = S_{\mathrm{ML}} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$$
(K^* is estimated ad hoc)

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

Laplace and KT

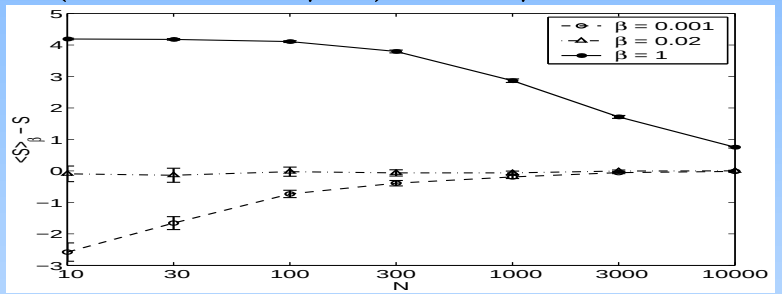
Problems of common estimators

Maximum likelihood

$$\mathcal{P}_0(S) = \delta(S)$$
$$S = S_{\mathrm{ML}} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$$
(K^* is estimated ad hoc)

Laplace and KT

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



Problems of common estimators

Maximum likelihood

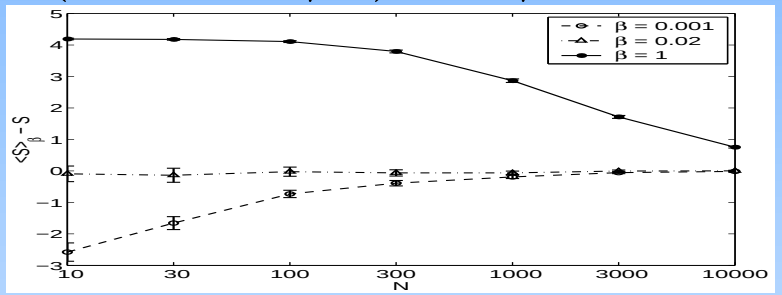
$$\mathcal{P}_0(S) = \delta(S)$$

$$S = S_{\mathrm{ML}} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$$

(K^* is estimated ad hoc)

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

Laplace and KT



Schurmann–Grassberger

Problems of common estimators

Maximum likelihood

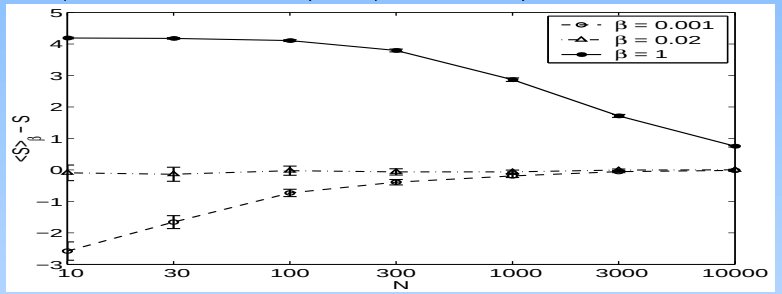
$\mathcal{P}_0(S) = \delta(S)$

$S = S_{\mathrm{ML}} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$

(K^* is estimated ad hoc)

Laplace and KT

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



Schurmann–Grassberger

$\sigma(1/K) \approx 0.61$ bit

(least biased)

Problems of common estimators

Maximum likelihood

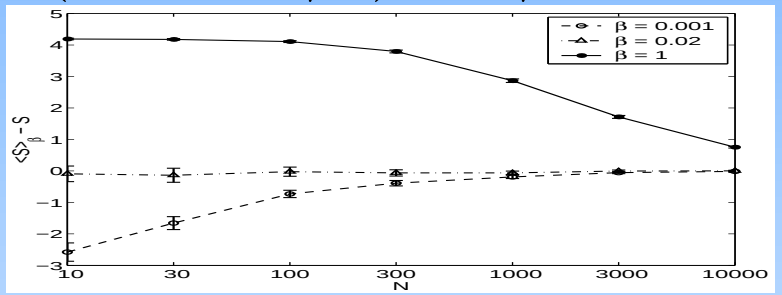
$\mathcal{P}_0(S) = \delta(S)$

$S = S_{\mathrm{ML}} + \frac{K^*}{2N} + O\left(\frac{1}{N^2}\right)$

(K^* is estimated ad hoc)

Laplace and KT

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



Schurmann–Grassberger

$\sigma(1/K) \approx 0.61$ bit

(least biased)

Still strongly biased towards $S = 1/\ln 2$ bits.

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

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

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

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

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

Solution: explanations

1. $d\xi/d\beta$ insures a priori uniformity over expected entropy.

Solution: explanations

1. $d\xi/d\beta$ insures a priori uniformity over expected entropy.
2. $\mathcal{P}(\xi)$ embodies actual expectations about entropy.

Solution: explanations

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.

Solution: explanations

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.
4. If $\rho(\xi)$ is peaked, then some $\beta(\xi)$ (model) dominates (is "selected"), and the variance of the estimator is small.

Too rough or too smooth?

Typical rank–ordered plots:

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

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

Too rough or too smooth?

Typical rank–ordered plots:

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

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

Faster decaying – too rough.
Slower decaying – too smooth.

Too rough or too smooth?

Typical rank–ordered plots:

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

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

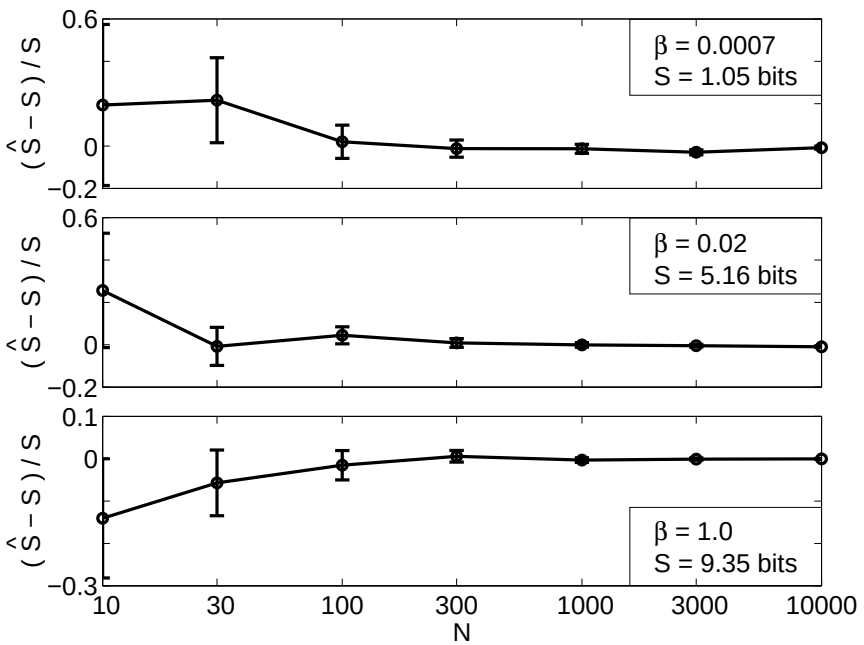
Faster decaying – too rough.

Slower decaying – too smooth.

Usually only the first regime is observed.

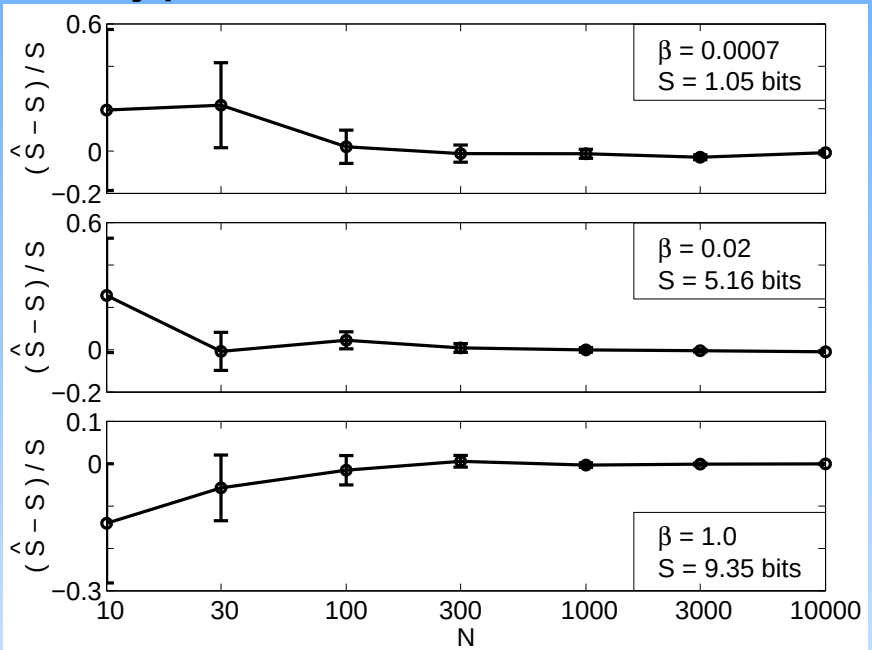
First attempts to estimate entropy

Typical distributions

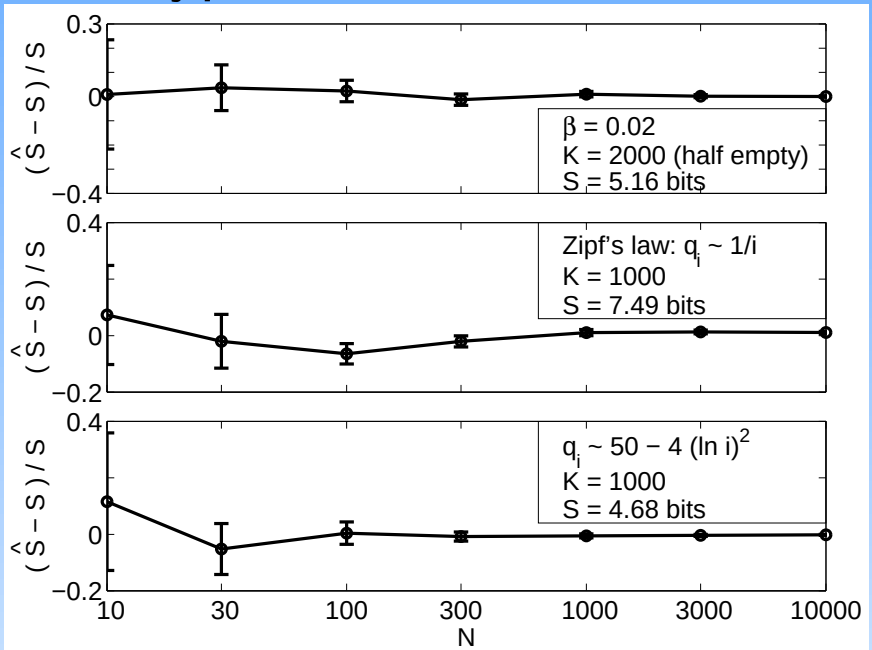


First attempts to estimate entropy

Typical distributions



Atypical distributions



Estimating entropy: first observations

1. Relative error $\sim 10\%$ at N as low as 30 for $K = 1000$.

Estimating entropy: first observations

1. Relative error $\sim 10\%$ at N as low as 30 for $K = 1000$.
2. Reliable estimation of error (posterior variance).

Estimating entropy: first observations

1. Relative error $\sim 10\%$ at N as low as 30 for $K = 1000$.
2. Reliable estimation of error (posterior variance).
3. *Little bias*, as it should be. Exception: too smooth distributions.

Estimating entropy: first observations

1. Relative error $\sim 10\%$ at N as low as 30 for $K = 1000$.
2. Reliable estimation of error (posterior variance).
3. *Little bias*, as it should be. Exception: too smooth distributions.
4. Key point: *learn entropies directly without finding* $\{q_i\}$*!*

Estimating entropy: first observations

1. Relative error $\sim 10\%$ at N as low as 30 for $K = 1000$.
2. Reliable estimation of error (posterior variance).
3. *Little bias*, as it should be. Exception: too smooth distributions.
4. Key point: *learn entropies directly without finding* $\{q_i\}$*!*
5. 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 – many coincidences

For $K \gg N \gg 1$, and $\Delta \equiv N - K$(nonzero counts) $\equiv N\delta \gg 1$, find $\beta^* = \kappa^*/N$ (saddle point).

$$\kappa^* = \kappa_0 + \frac{1}{K}\kappa_1 + \frac{1}{K^2}\kappa_2 + \dots$$

Asymptotics – many coincidences

For $K \gg N \gg 1$, and $\Delta \equiv N - K$(nonzero counts) $\equiv N\delta \gg 1$, find $\beta^* = \kappa^*/N$ (saddle point).

$$\begin{aligned} \kappa^* &= \kappa_0 + \frac{1}{K}\kappa_1 + \frac{1}{K^2}\kappa_2 + \dots \\ \kappa_0 &= N\left(\frac{b_{-1}}{\delta} + b_0 + b_1\delta + \dots\right) \end{aligned}$$

Asymptotics – many coincidences

For $K \gg N \gg 1$, and $\Delta \equiv N - K$(nonzero counts) $\equiv N\delta \gg 1$, find $\beta^* = \kappa^*/N$ (saddle point).

$$\begin{aligned} \kappa^* &= \kappa_0 + \frac{1}{K}\kappa_1 + \frac{1}{K^2}\kappa_2 + \dots \\ \kappa_0 &= N\left(\frac{b_{-1}}{\delta} + b_0 + b_1\delta + \dots\right) \end{aligned}$$

other κ_i and b_i are $O(1)$

Asymptotics – many coincidences

For $K \gg N \gg 1$, and $\Delta \equiv N - K$(nonzero counts) $\equiv N\delta \gg 1$, find $\beta^* = \kappa^*/N$ (saddle point).

$$\begin{aligned} \kappa^* &= \kappa_0 + \frac{1}{K}\kappa_1 + \frac{1}{K^2}\kappa_2 + \dots \\ \kappa_0 &= N\left(\frac{b_{-1}}{\delta} + b_0 + b_1\delta + \dots\right) \\ & \quad \text{other } \kappa_i \text{ and } b_i \text{ are } O(1) \end{aligned}$$

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

Asymptotics – few coincidences

For $K \to \infty$, $\Delta \sim 1$, $\delta \to 0$

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

Estimator: Properties

1. K can potentially be infinite.

Estimator: Properties

1. K can potentially be infinite.
2. Estimation for small Δ is only reliable if distribution is not atypically smooth.

Estimator: Properties

1. K can potentially be infinite.
2. Estimation for small Δ is only reliable if distribution is not atypically smooth.
3. Expansion parameter for saddle point analysis is Δ.

Estimator: Properties

1. K can potentially be infinite.
2. Estimation for small Δ is only reliable if distribution is not atypically smooth.
3. Expansion parameter for saddle point analysis is Δ.
4. Selection of K by Bayesian integration not an option: small K means smaller phase space and *better* approximation.

Estimator: Properties

1. K can potentially be infinite.
2. Estimation for small Δ is only reliable if distribution is not atypically smooth.
3. Expansion parameter for saddle point analysis is Δ.
4. Selection of K by Bayesian integration not an option: small K means smaller phase space and *better* approximation.
5. The estimator is consistent.

Estimator: Properties

1. K can potentially be infinite.
2. Estimation for small Δ is only reliable if distribution is not atypically smooth.
3. Expansion parameter for saddle point analysis is Δ.
4. Selection of K by Bayesian integration not an option: small K means smaller phase space and *better* approximation.
5. The estimator is consistent.
6. The estimator should work (in some cases) for $N \ll K$, $N \ll 2^S$, and $N \sim 2^{S/2}$ (cf. Ma, 1981).

Estimator: Synthetic test

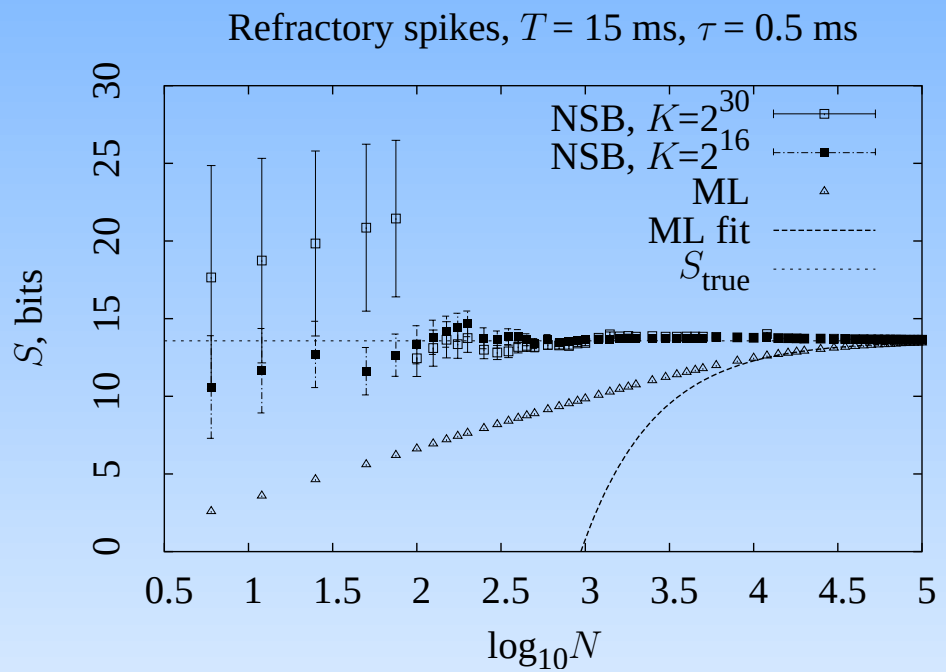
Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$.

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

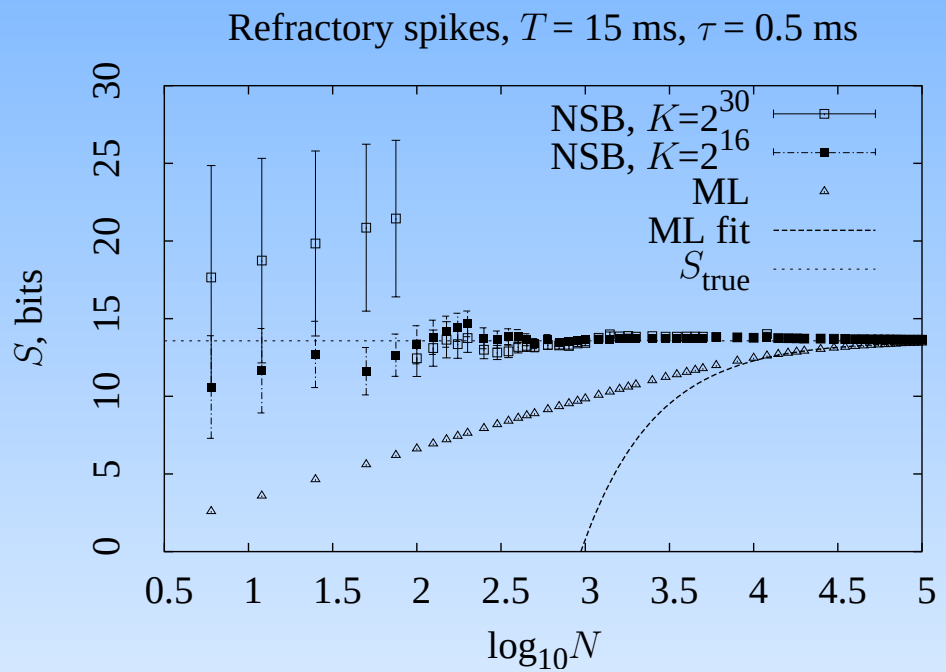
Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$. $K = 2^{30}$, $K_{\text{ref}} < 2^{16}$, $S = 13.57\text{bits}$.



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

Estimator: Synthetic test

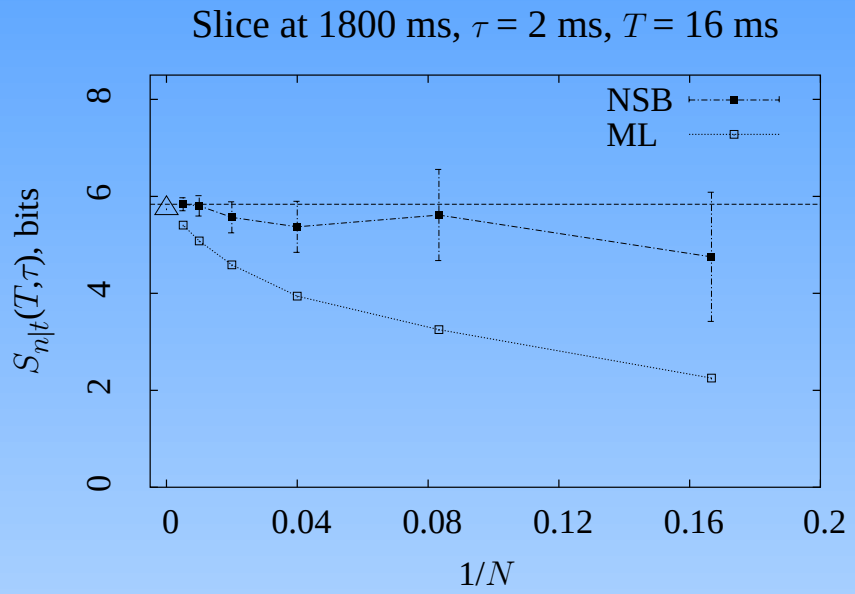
Refractory Poisson process: $r = 0.26\text{ms}^{-1}$, $R = 1.8\text{ms}$, $T = 15\text{ms}$, $\tau = 0.5\text{ms}$. $K = 2^{30}$, $K_{\text{ref}} < 2^{16}$, $S = 13.57\text{bits}$.



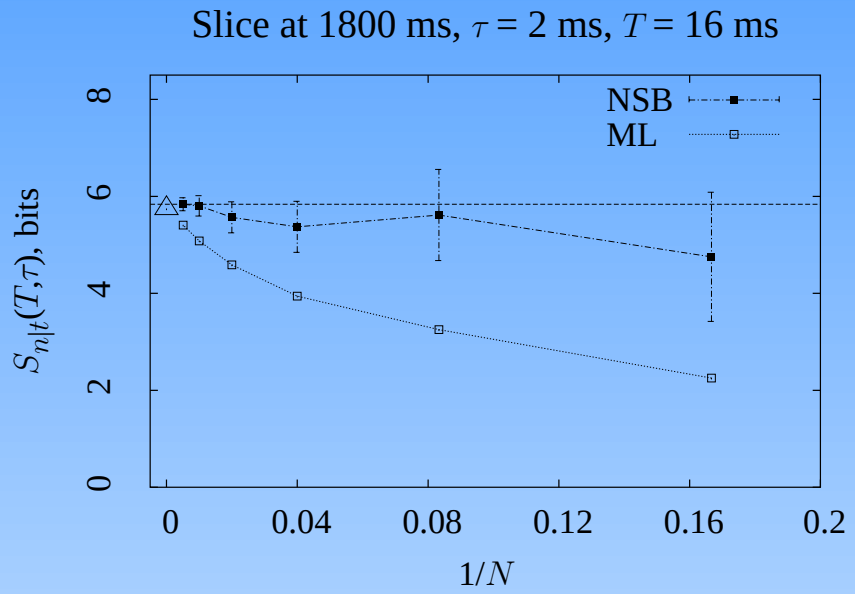
True value reached within the error bars for $N^2 \sim 2^S$, when coincidences start to occur.

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

Natural data: Slice entropy vs. sample size



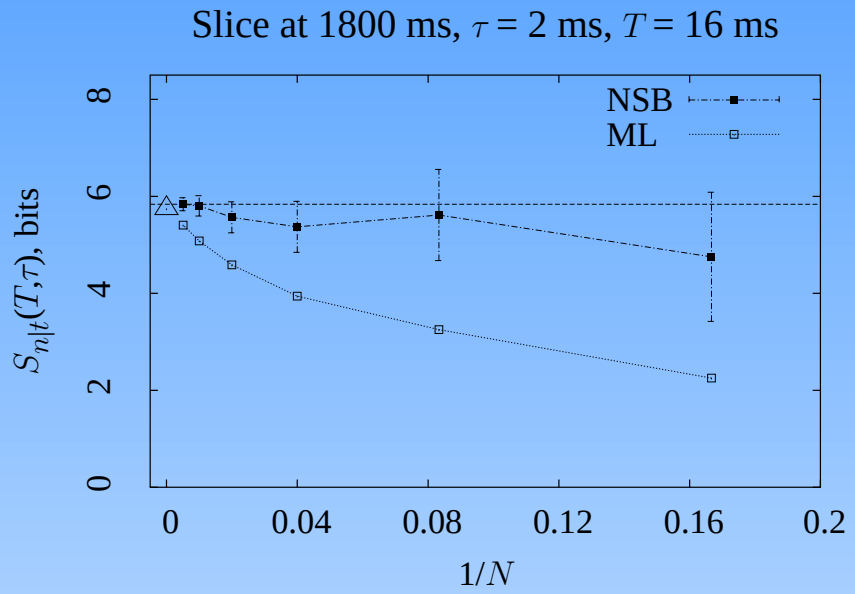
Natural data: Slice entropy vs. sample size



ML estimator converges with $\sim 1/N$

corrections.

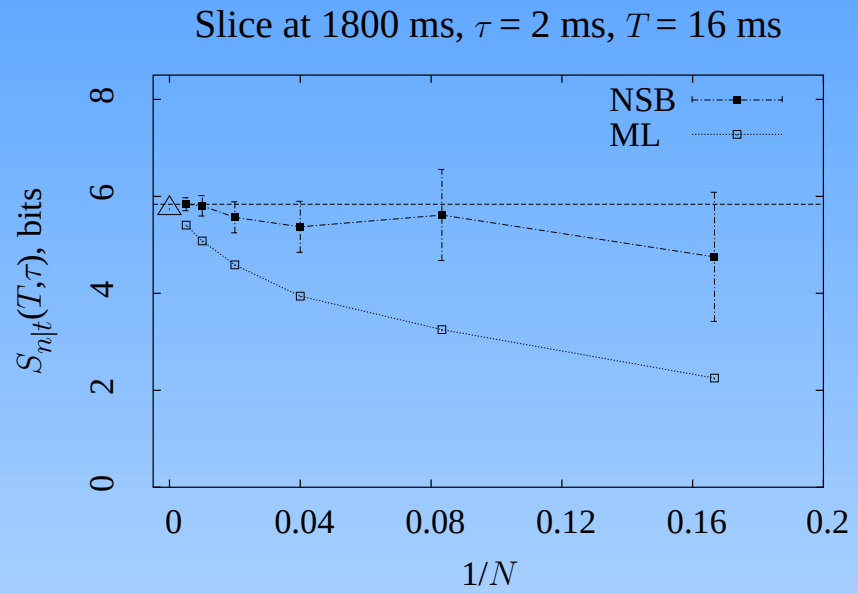
Natural data: Slice entropy vs. sample size

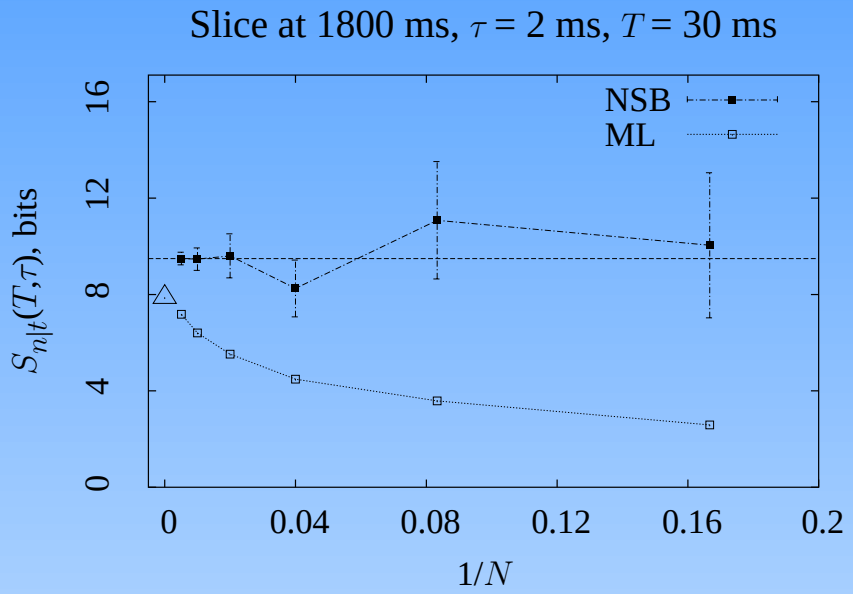


ML estimator converges with $\sim 1/N$ corrections.

NSB estimator is always within error bars.

Natural data: Slice entropy vs. sample size

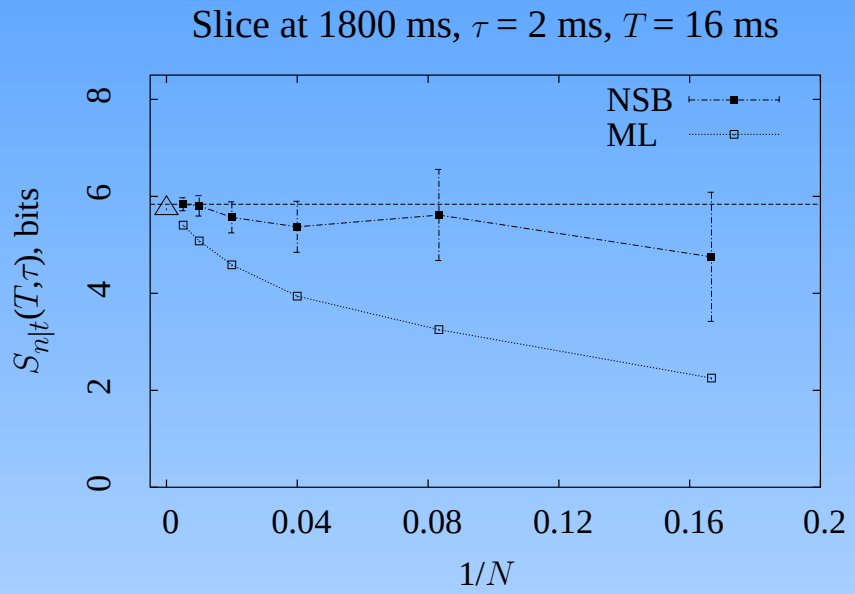


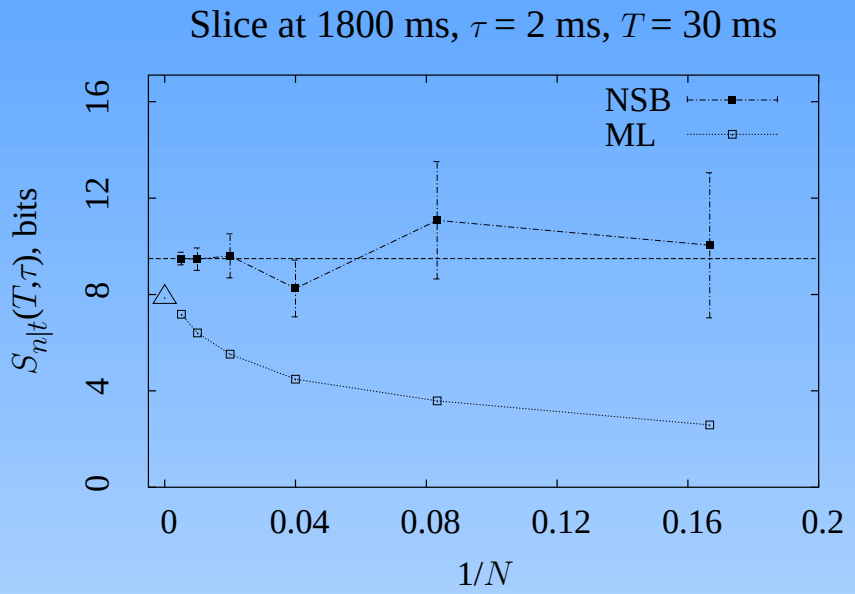


ML estimator converges with $\sim 1/N$ corrections.

NSB estimator is always within error bars.

Natural data: Slice entropy vs. sample size



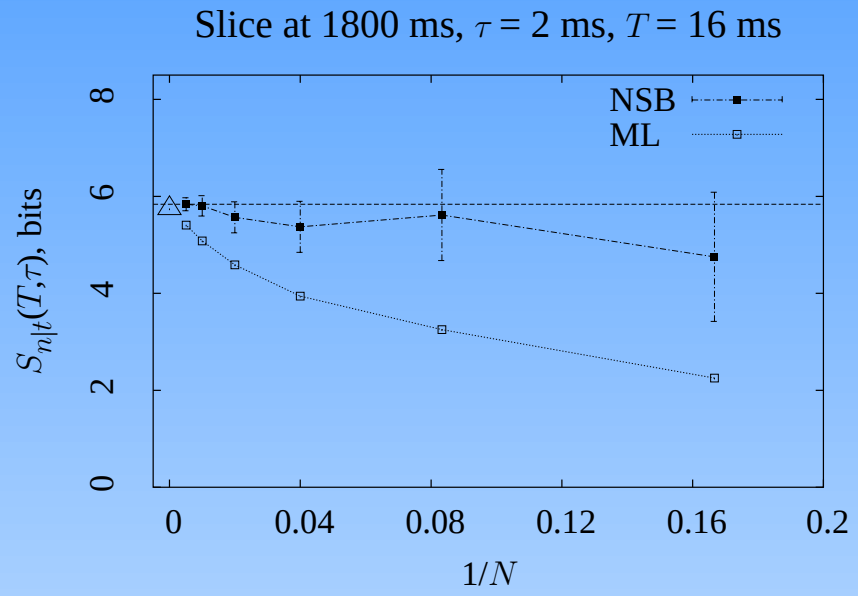


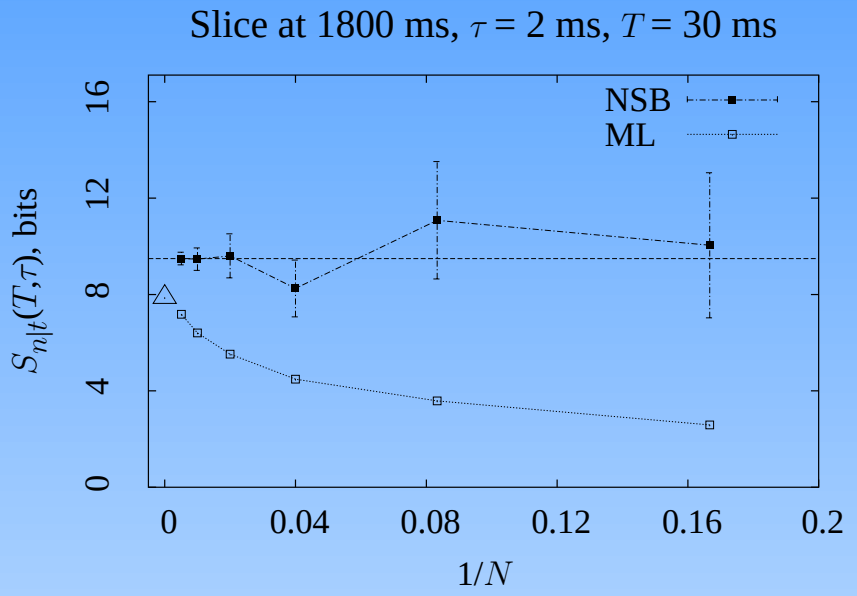
ML estimator converges with $\sim 1/N$ corrections.

NSB estimator is always within error bars.

ML estimator cannot be extrapolated.

Natural data: Slice entropy vs. sample size





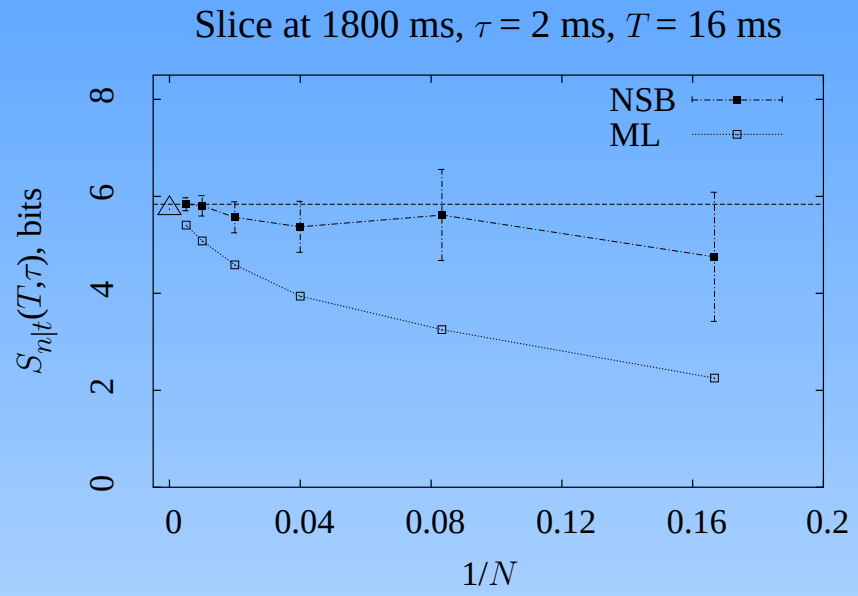
ML estimator converges with $\sim 1/N$ corrections.

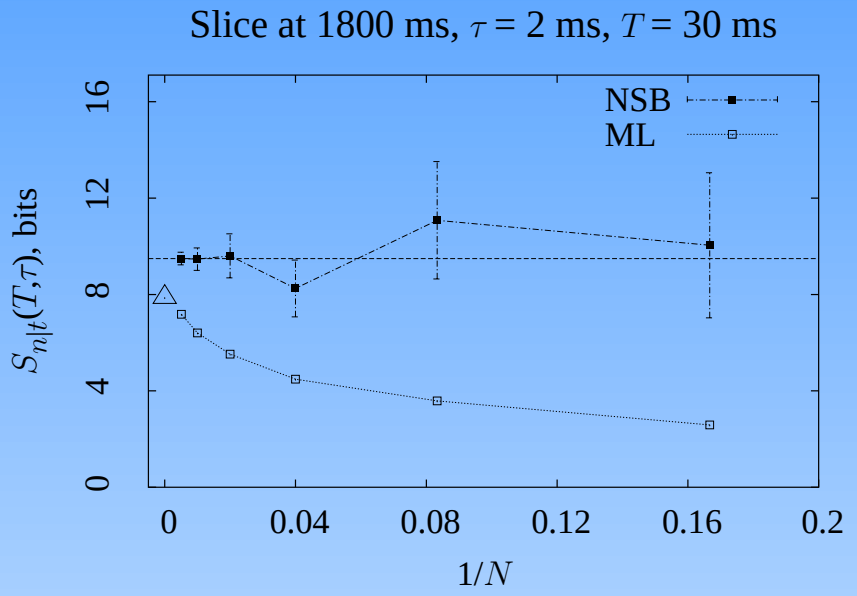
NSB estimator is always within error bars.

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.

$(S^{\mathrm{NSB}} - S_{\mathrm{ML}})/\delta S^{\mathrm{NSB}}$ has zero mean if S^{ML} is reliable.

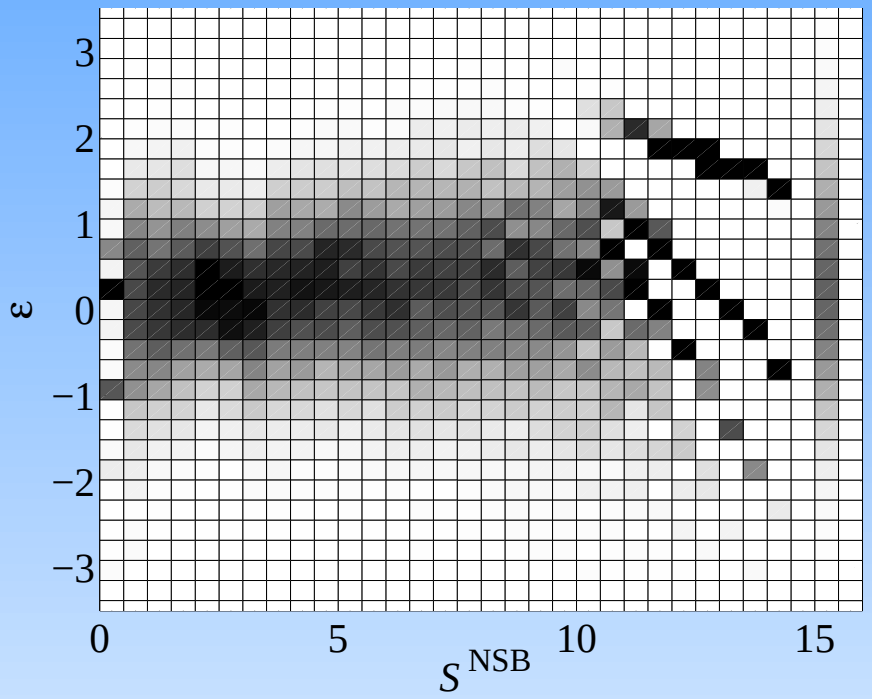
Natural data: Error vs. mean

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

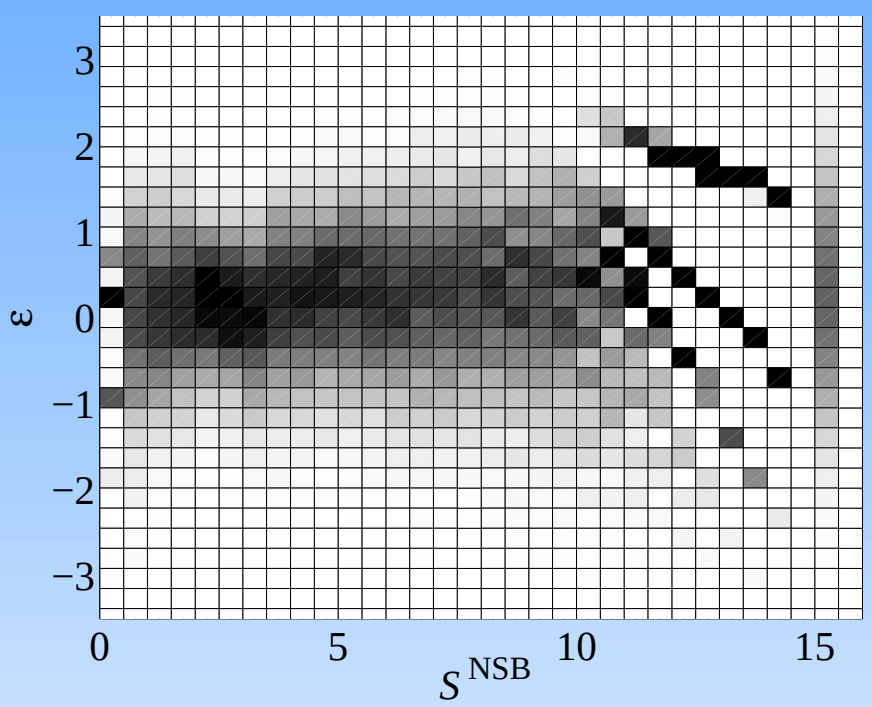
$N = 75$



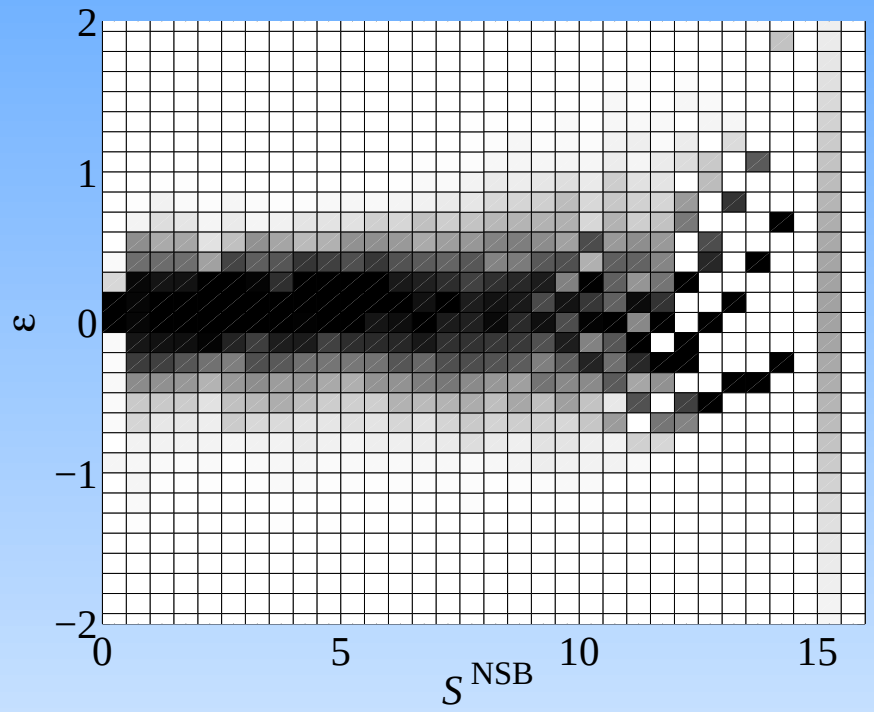
Natural data: Error vs. mean

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

$N = 75$



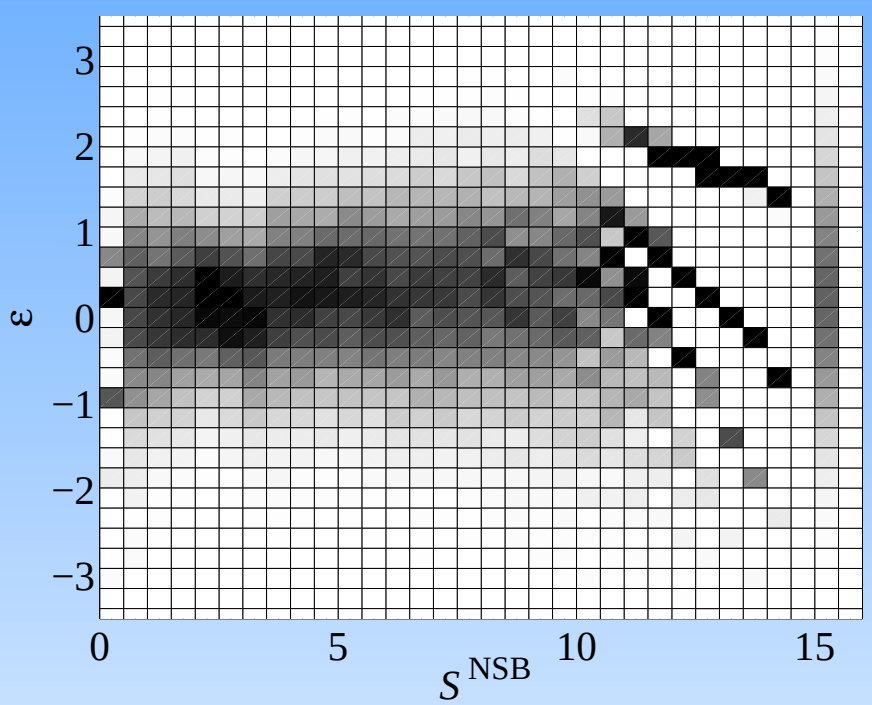
$N = 175$



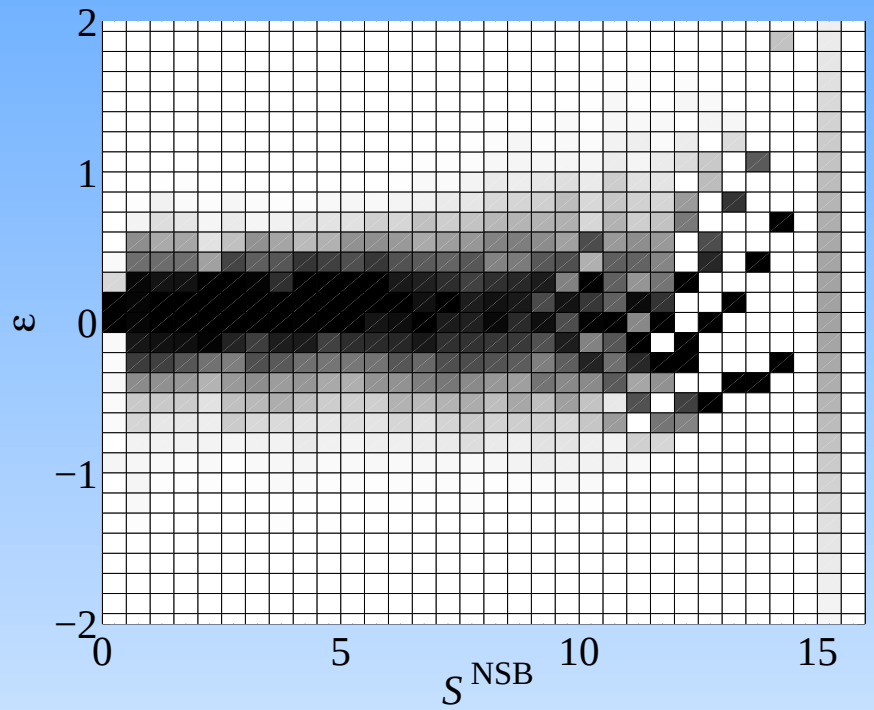
Natural data: Error vs. mean

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

$N = 75$



$N = 175$

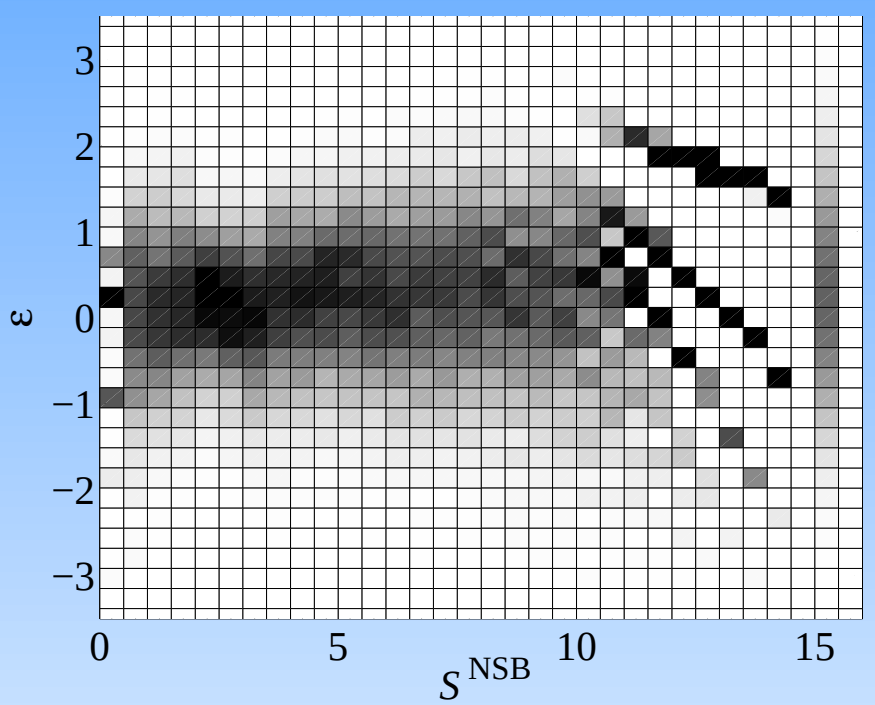


Almost no bias.

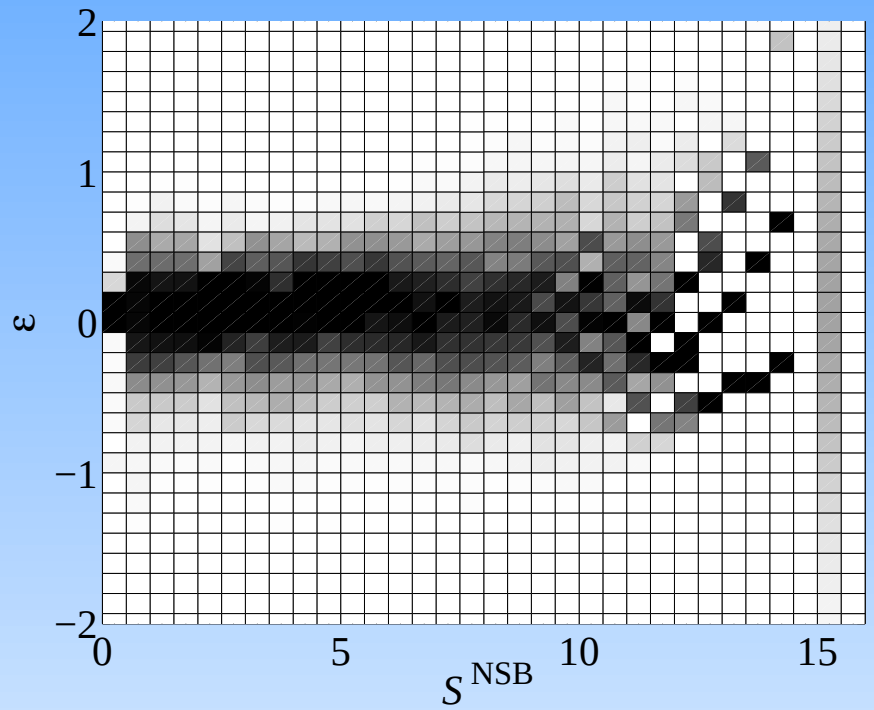
Natural data: Error vs. mean

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

$N = 75$



$N = 175$

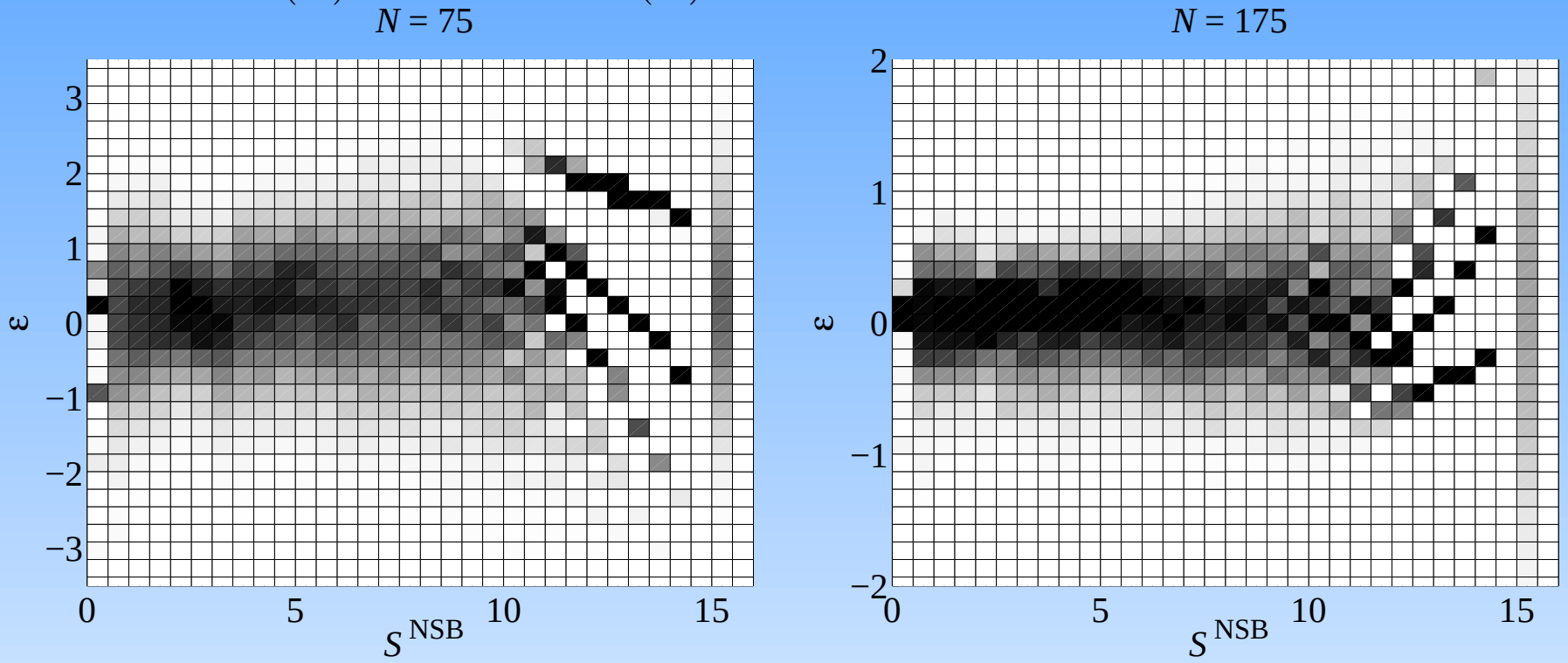


Almost no bias.

Empirical variance < 1 due to long tails in posterior.

Natural data: Error vs. mean

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



Almost no bias.

Empirical variance < 1 due to long tails in posterior.

Bands are due to discrete nature of Δ.

Natural data: Hints of future results

Some preliminary results for information rate estimation. Further work is needed to properly estimate error bars due to signal correlations.

Natural data: Hints of future results

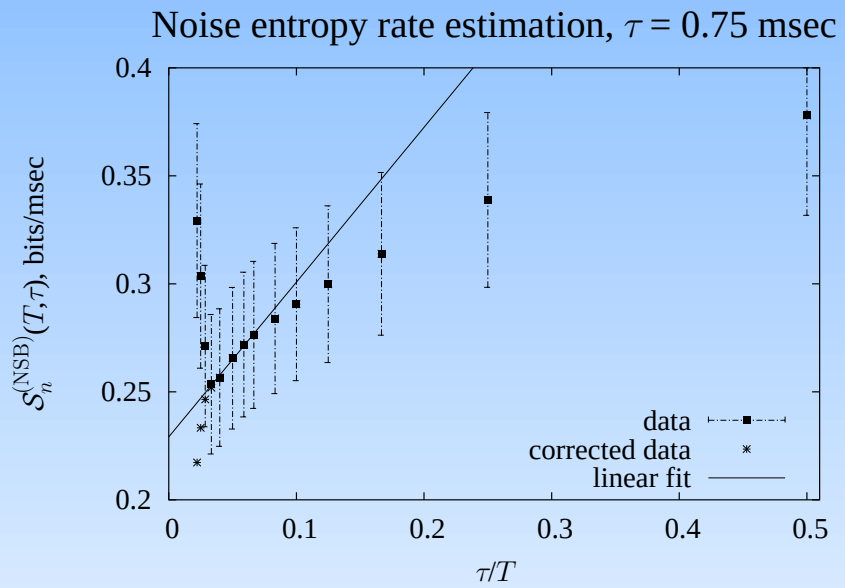
Some preliminary results for information rate estimation. Further work is needed to properly estimate error bars due to signal correlations.

The fly in question is nosier than usual.

Natural data: Hints of future results

Some preliminary results for information rate estimation. Further work is needed to properly estimate error bars due to signal correlations.

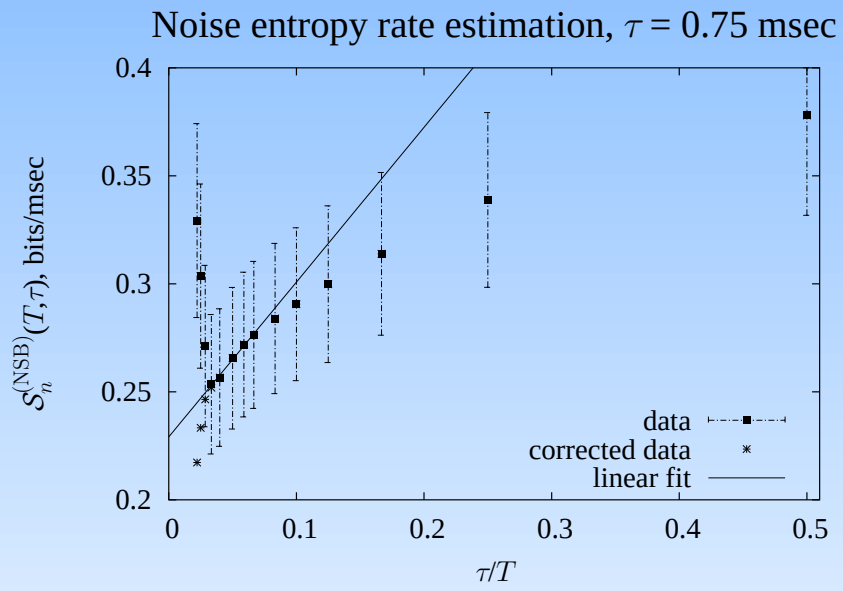
The fly in question is nosier than usual.

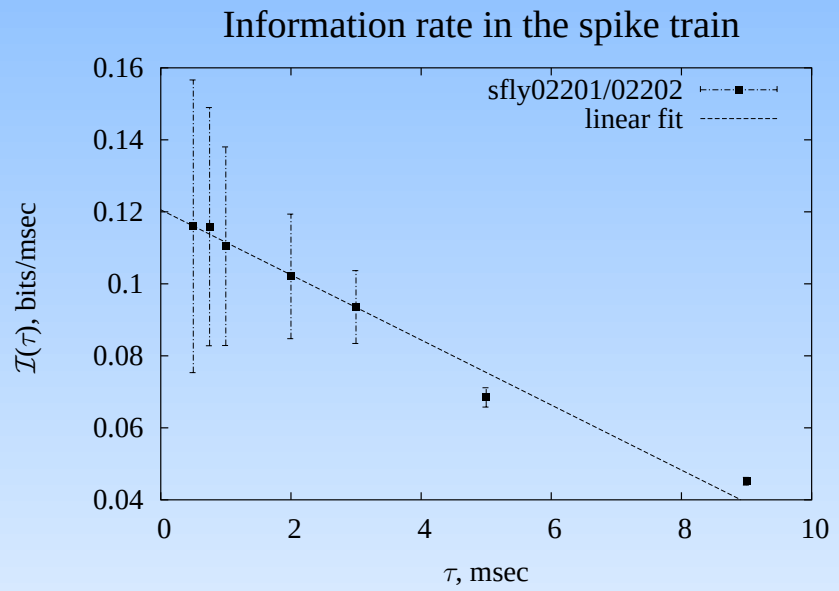


Natural data: Hints of future results

Some preliminary results for information rate estimation. Further work is needed to properly estimate error bars due to signal correlations.

The fly in question is nosier than usual.





Conclusions

1. Found new entropy estimator.

Conclusions

1. Found new entropy estimator.

2. Know if we should trust it.

Conclusions

1. Found new entropy estimator.
2. Know if we should trust it.
3. Neural data seems to be well matched to the estimator.