Estimation of information-theoretic quantities

Liam Paninski

Gatsby Computational Neuroscience Unit
University College London
http://www.gatsby.ucl.ac.uk/∼liam
liam@gatsby.ucl.ac.uk

November 16, 2004

Estimation of information

Some questions:

- What part of the sensory input is best encoded by a given neuron?
- Are early sensory systems optimized to transmit information?
- Do noisy synapses limit the rate of information flow from neuron to neuron?

Need to quantify "information."

Mutual information

$$I(X;Y) = \int_{\mathcal{X}\times\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mathematical reasons:

- invariance
- "uncertainty" axioms
- data processing inequality
- channel and source coding theorems

But obvious open experimental question:

- is this computable for real data?

How to estimate information

I very hard to estimate in general...

... but lower bounds (via data processing inequality) are easier.

Two ideas:

1) decoding approach: estimate $x|y$, use quality of estimate to lower bound $I(X;Y)$

2) discretize x, y, estimate discrete information
$I_{discrete}(X;Y) \leq I(X;Y)$

Decoding lower bound

$$
\begin{aligned}
I(X;Y) &\geq I(X;\hat{X}(Y)) && (1)\\
&= H(X) - H(X|\hat{X}(Y)) \\
&\geq H(X) - H[\mathcal{N}(0, \mathrm{Cov}(X|\hat{X}(Y))] && (2)
\end{aligned}
$$

(1): Data processing inequality

(2): Gaussian maxent: $H[\mathcal{N}(0, \mathrm{Cov}(X|\hat{X}(Y))] \geq H(X|\hat{X}(Y))$

(Rieke et al., 1997)

Gaussian stimuli

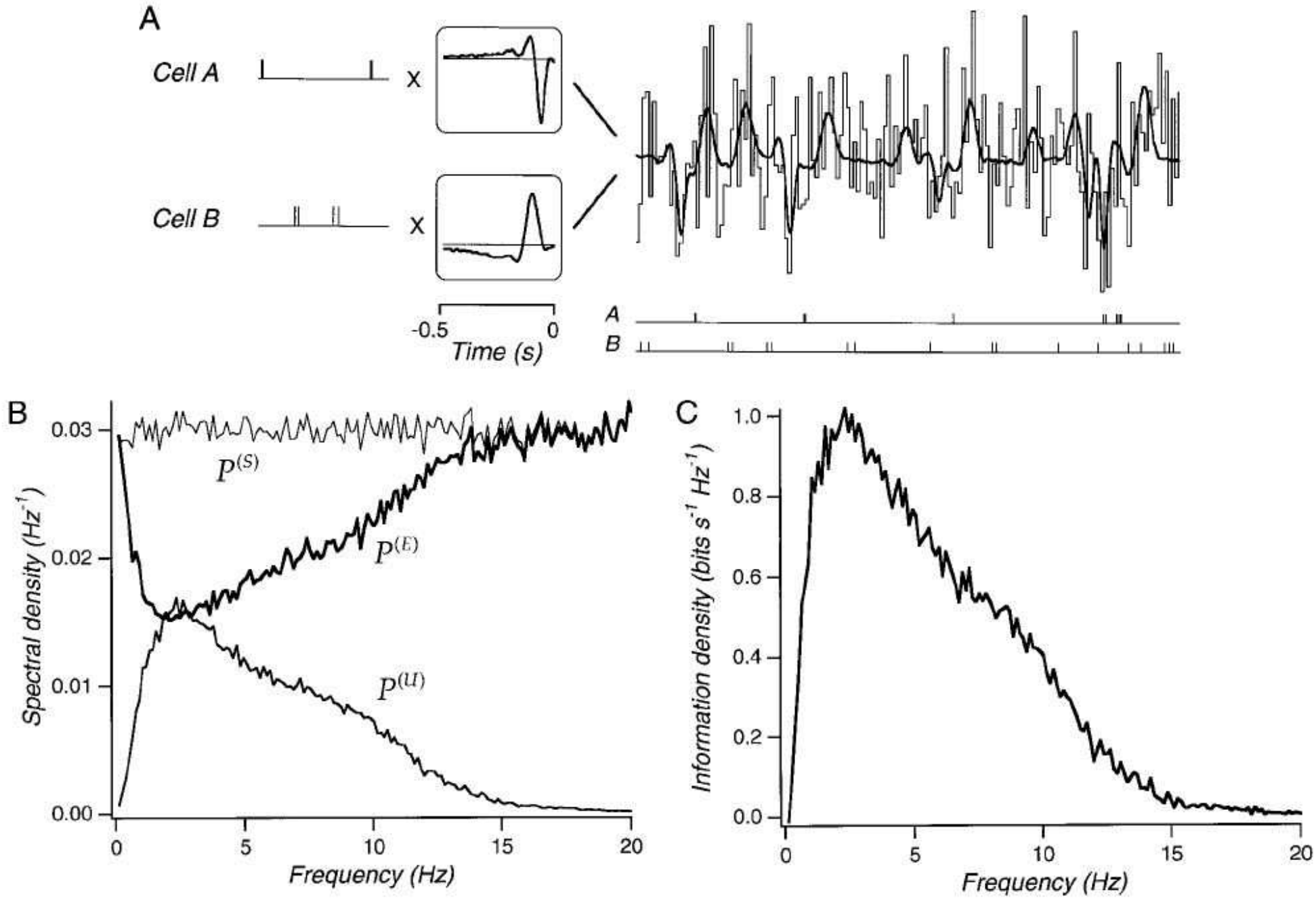
$X(t)$ Gaussian + stationary $\implies$ specified by power spectrum
(= covariance in Fourier domain)

Use Shannon formula $\dot{I} = \int d\omega \log SNR(\omega)$

$SNR(\omega)$ = signal-to-noise ratio at frequency ω

(Rieke et al., 1997)

Calculating the noise spectrum



(Warland et al., 1997)

Pros and cons

Pros:

— only need to estimate covariances, not full distribution

— Fourier analysis gives good picture of what information is kept, discarded

Cons:

— tightness of lower bound depends on decoder quality

— bound can be inaccurate if noise is non-Gaussian

Can we estimate I without a decoder, for general X, Y?

Discretization approach

Second approach: discretize spike train into one of m bins, estimate

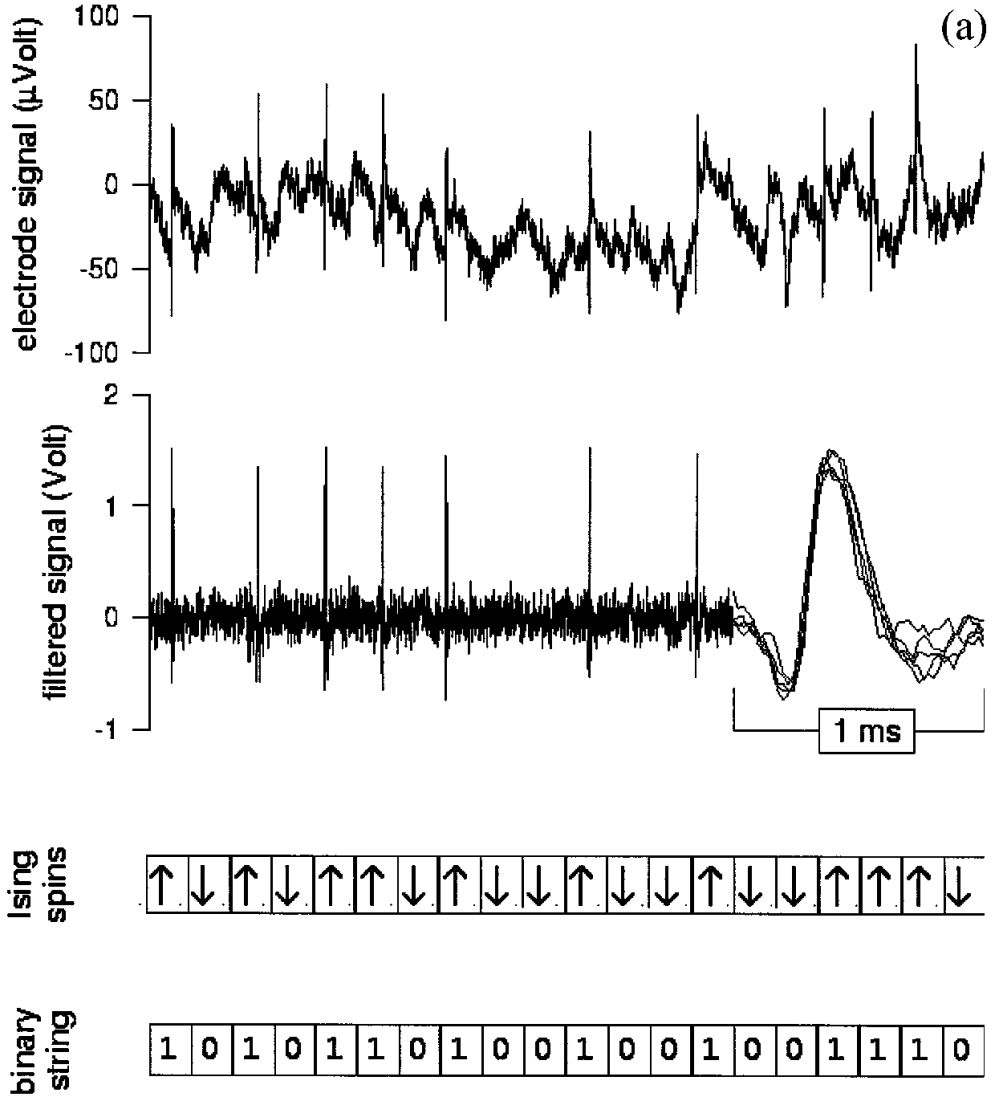
$$I_{discrete}(X;Y) = \sum_m p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Data processing: $I_{discrete}(X;Y) \leq I(X;Y)$, for any m.

Refine as more data come in; if m grows slowly enough, $\hat{I}_{discrete} \to I_{discrete} \nearrow I$.

— doesn't assume *anything* about X or code $p(y|x)$: as nonparametric as possible

Digitizing spike trains



To compute entropy *rate*, take limit $T \to \infty$ (Strong et al., 1998)

Discretization approach

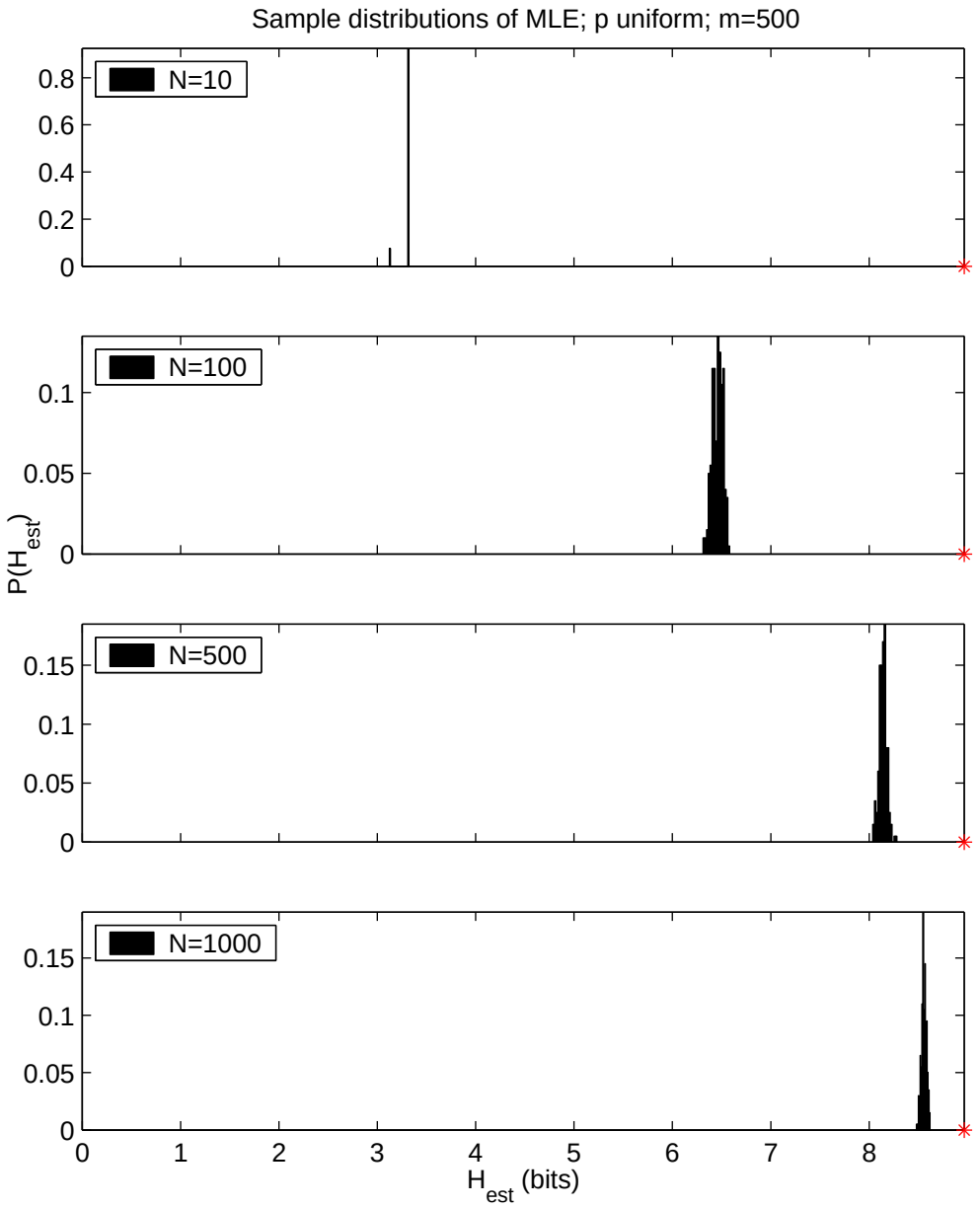
Use MLE to estimate H (if we have H, we have I):

$$\hat{H}_{MLE}(p_n) \equiv -\sum_{i=1}^{m} p_n(i) \log p_n(i)$$

Obvious concerns:

- Want $N >> m$ samples, to "fill in" histograms $p(x, y)$
- How large is bias?

Bias is major problem



N = number of samples

Bias is major problem

- $\hat{H}_{MLE}$ is negatively biased for all p
- Rough estimate of $B(\hat{H}_{MLE})$: $-(m-1)/2N$.
- Variance is much smaller: $\sim (\log m)^2/N$
- No unbiased estimator exists

(Exercise: prove each of the above statements.)

Try "bias-corrected" estimator:

$$\hat{H}_{MM} \equiv \hat{H}_{MLE} + \frac{\hat{m}-1}{2N}$$

— $\hat{H}_{MM}$ due to (Miller, 1955); see also e.g. (Treves and Panzeri, 1995)

Convergence of common information estimators

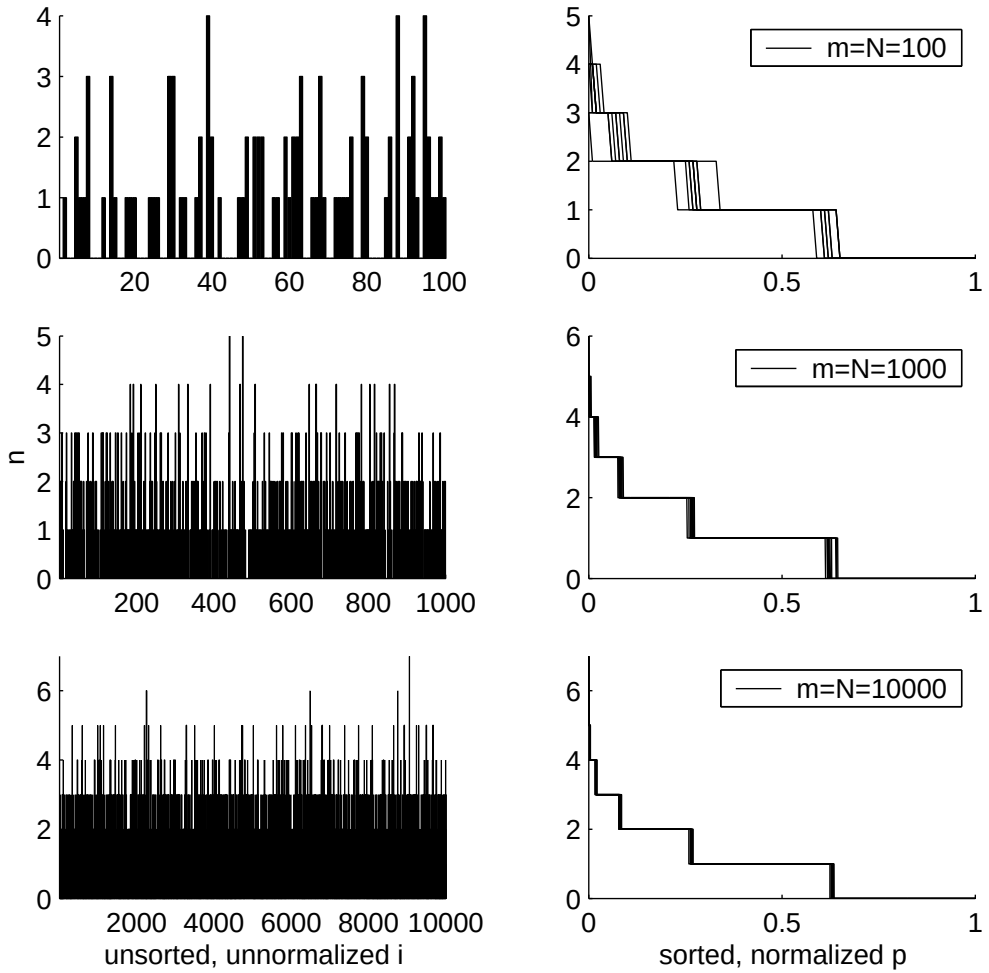
Result 1: If $N/m \to \infty$, ML and bias-corrected estimators converge to right answer.

Converse: if $N/m \to c < \infty$, ML and bias-corrected estimators converge to *wrong* answer.

Implication: if N/m small, bias is large although errorbars vanish — *even if "bias-corrected" estimators are used* (Paninski, 2003)

Whence bias?

Sample histograms from uniform density:



If $N/m \to c < \infty$, sorted histograms converge to wrong density.

Variability in histograms $\implies$ bias in entropy estimates

Estimating information on m bins with fewer than m samples

Result 2: A new estimator that gives correct answer even if $N < m$

— Estimator works well even in *worst* case

Interpretation: entropy is easier to estimate than p!

$\Longrightarrow$ we can estimate the information carried by the neural code even in cases when the codebook $p(y|x)$ is too complex to learn directly (Paninski, 2003; Paninski, 2004).

Sketch of logic

- Good estimators have low error for all p
- Error is sum of bias and variance

Goal:

1. find simple "worst-case" bounds on bias, variance
2. minimize bounds over some large but tractable class of estimators

A simple class of estimators

- Entropy is $\sum f(n_i)$, $f(t) = -\frac{t}{N} \log \frac{t}{N}$.
- Look for $\hat{H} = \sum g_{N,m}(n_i)$, where $g_{N,m}$ minimizes worst-case error
- $g_{N,m}$ is just an $(N+1)$-vector
- Very simple class, but turns out to be rich enough

Deriving a bias bound

$$
\begin{aligned}
B &= E(\hat{H}) - H \\
&= E(\sum_i g(n_i)) - \sum_i f(p_i) \\
&= \sum_j \sum_i P(n_i = j) g(n_i) - \sum_i f(p_i) \\
&= \sum_i \left(\sum_j B_j(p_i) g(j) \right) - f(p_i)
\end{aligned}
$$

- $B_j(p) = \binom{N}{j} p^j (1-p)^{N-j}$: polynomial in p
- If $\sum_j g(j) B_j(p)$ close to $f(p)$ for all p, bias will be small
- Standard uniform polynomial approximation theory

Bias and variance

- Interesting point: can make bias very small ($\sim m/N^2$), but variance explodes, ruining estimator.
- In fact, no uniform bounds can hold if $m > N^\alpha, \alpha > 1$
- Have to bound bias and variance together

Variance bound

"Method of bounded differences"

$F(x_1, x_2, ..., x_N)$ a function of N i.i.d. r.v.'s

If any single x_i has small effect on F, i.e,

$\max |F(..., x, ...) - F(..., y, ...)|$ small

$\implies$ $var(F)$ small.

Our case: $\hat{H} = \sum_i g(n_i)$;

$\max_j |g(j) - g(j-1)|$ small $\implies$ $\mathrm{Var}(\sum_i g(n_i))$ small

Computation

Goal: minimize $\max_p$ (bias^2 + variance)

- bias $\leq m \cdot \max_{0 \leq t \leq 1} |f(t) - \sum_j g(j) B_j(t)|$
- variance $\leq N \max_j |g(j) - g(j-1)|^2$

Idea: minimize sum of bounds

Convex in $\vec{g} \implies$ tractable

...but still slow for N large

Fast solution

Trick 1: approximate maximum error by mean-square error

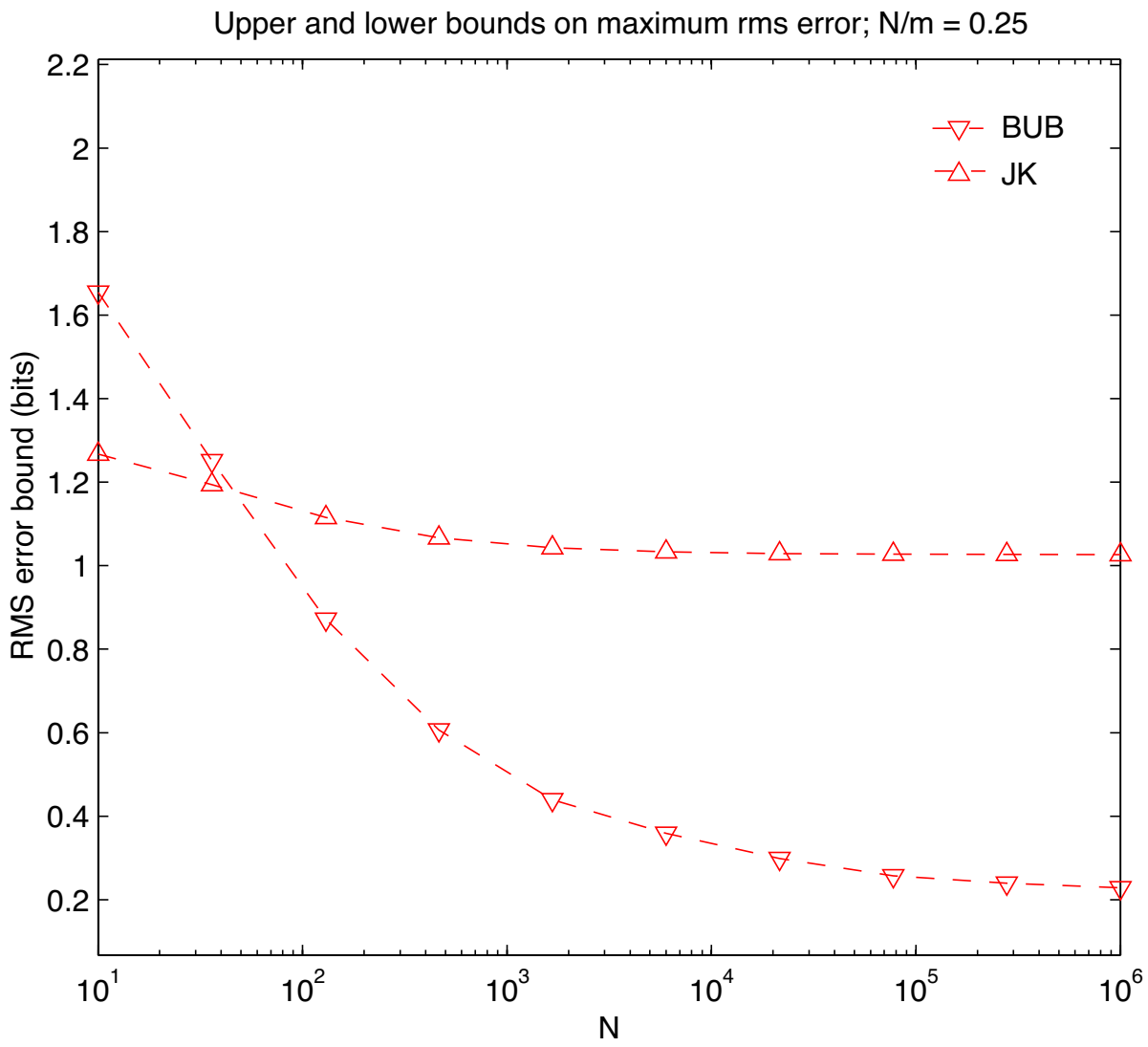
$\Longrightarrow$ simple regression problem: good solution in $O(N^3)$ time

Trick 2: use good starting point from approximation theory

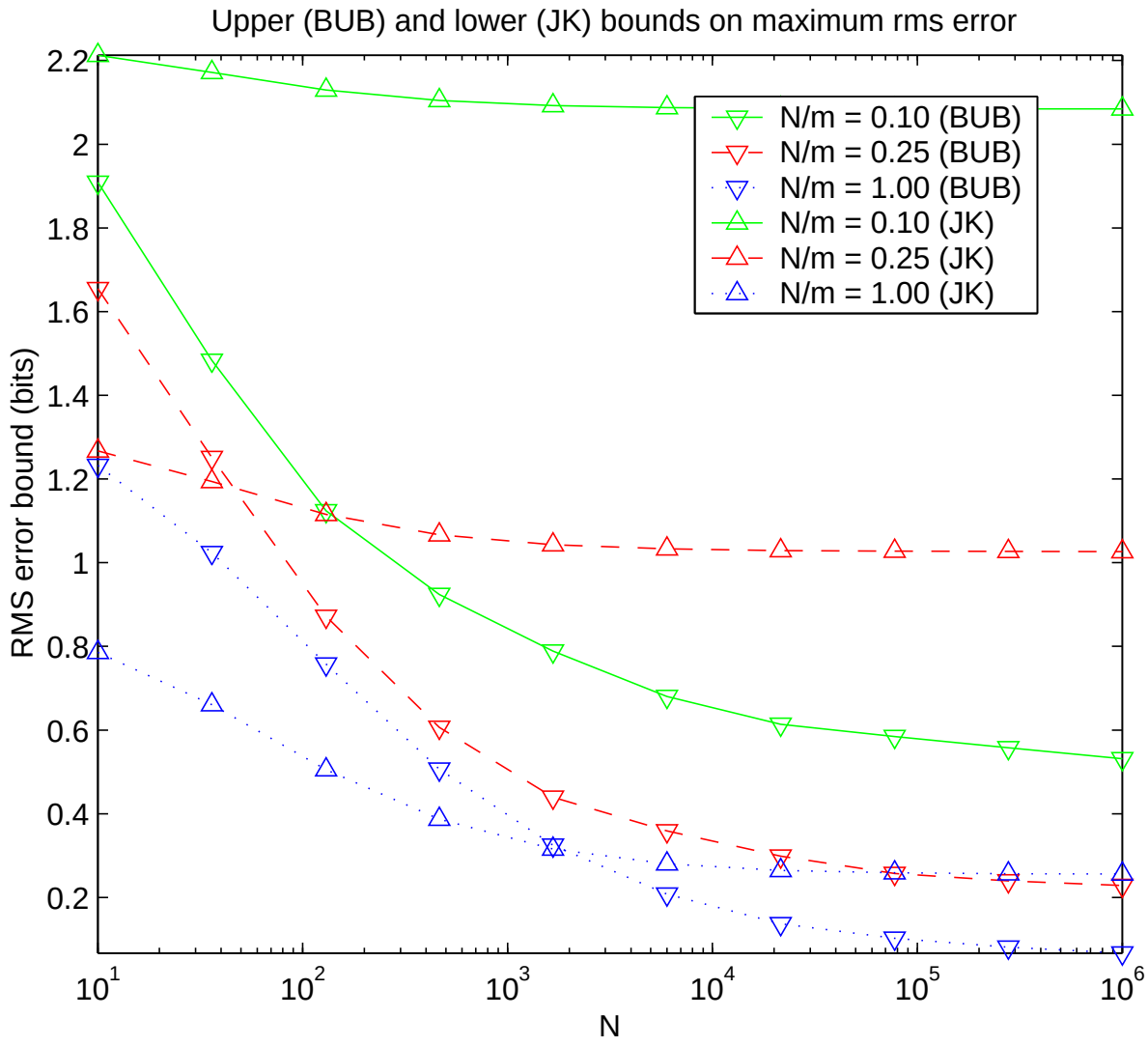
$\Longrightarrow$ good solution in $O(N)$ time

- Computation time independent of m
- Once $g_{N,m}$ is calculated, cost is exactly same as for $p \log p$

Error comparisons: upper and lower bounds

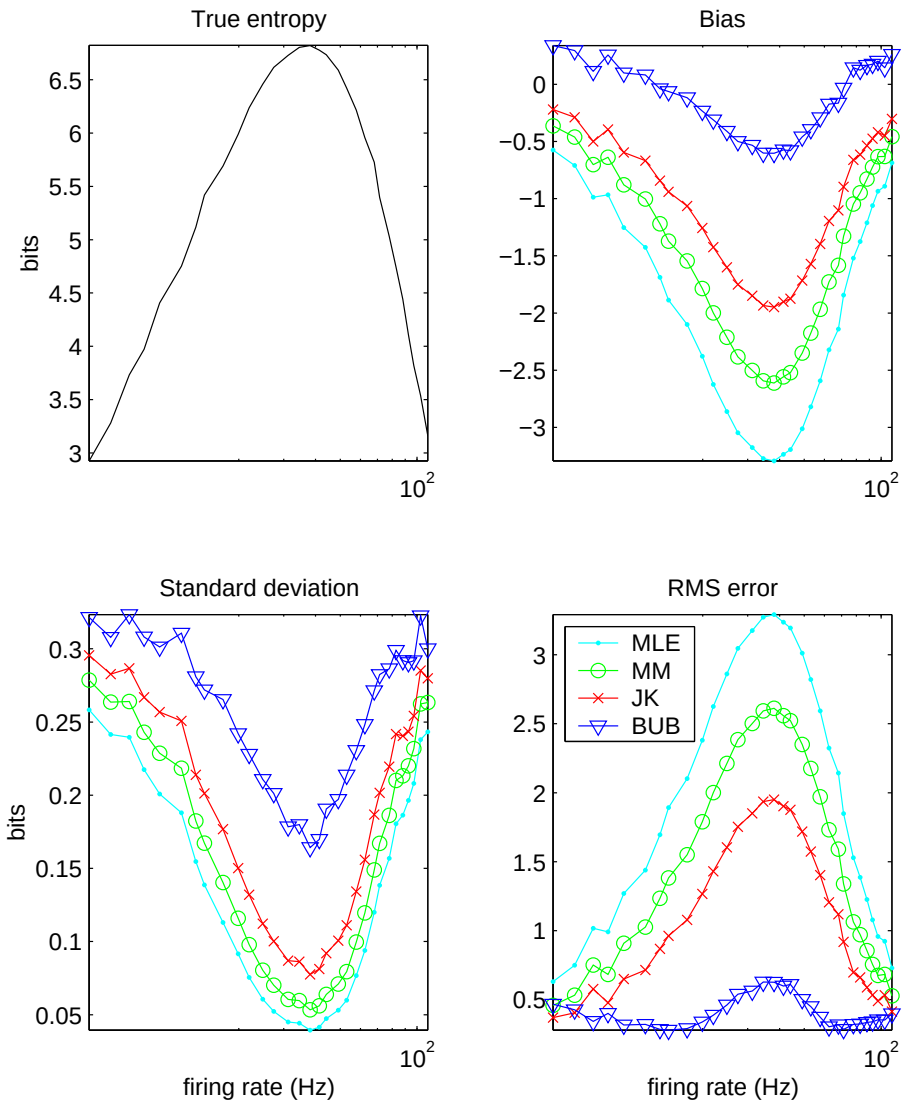


Error comparisons: upper and lower bounds



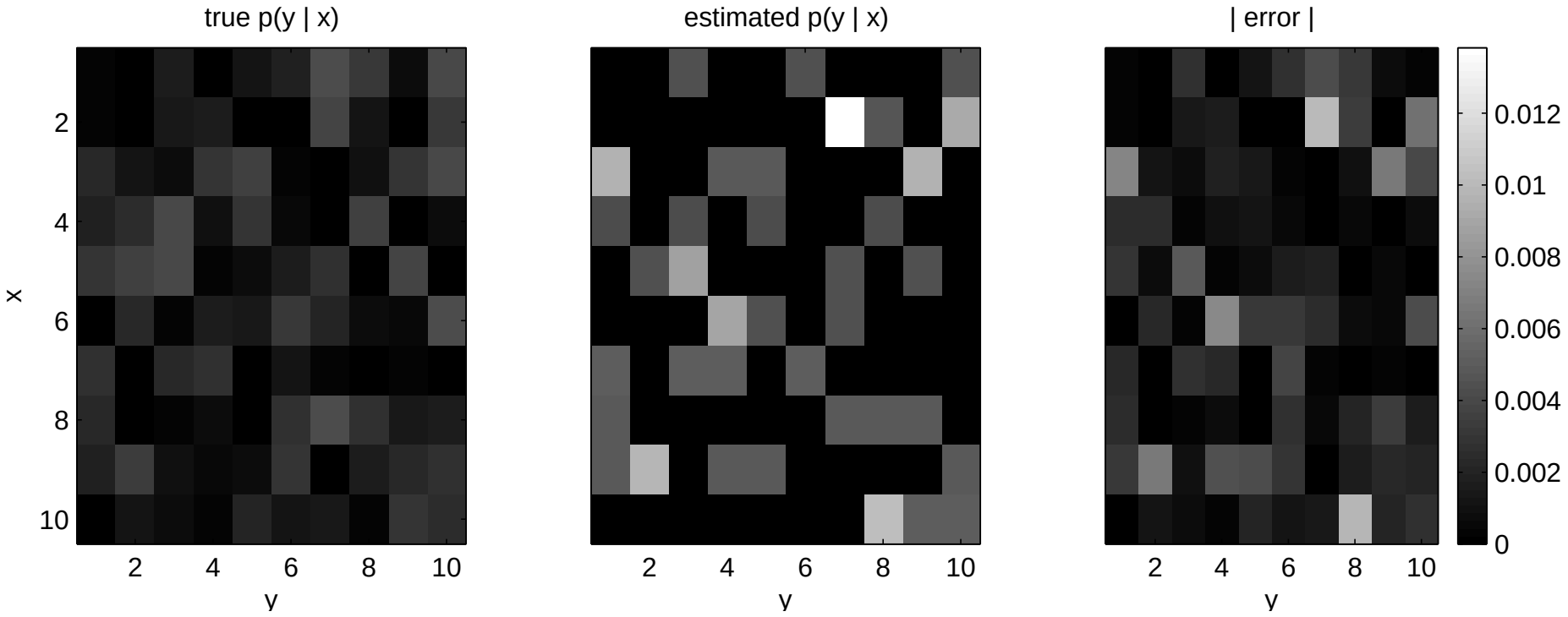
— upper bound on worst-case error $\to 0$, even if $N/m > c > 0$.

Error comparisons: integrate-and-fire model data



Similar effects both *in vivo* and *in vitro*

Undersampling example



$m_x = m_y = 700; N/m_{xy} = 0.3$

$\hat{I}_{MLE} = 2.21$ bits

$\hat{I}_{MM} = -0.19$ bits

$\hat{I}_{BUB} = \mathbf{0.60}$ bits; conservative (worst-case upper bound) error: ± 0.2 bits

true $I(X;Y) = \mathbf{0.62}$ bits

Other approaches

— Compression

— Bayesian estimators

— Parametric modeling

Compression approaches

Use interpretation of entropy as total number of bits required to code signal ("source coding" theorem)

Apply a compression algorithm (e.g. Lempel-Ziv) to data, take $\hat{H}$ = number of bits required

— takes temporal nature of data into account more directly than discretization approach

Bayesian approaches

Previous analysis was "worst-case": applicable without any knowledge at all of the underlying p.

Easy to perform Bayesian inference if we have *a priori* knowledge of

— p (Wolpert and Wolf, 1995)

— $H(p)$ (Nemenman et al., 2002)

(Note: "ignorant" priors on p can place very strong constraints on $H(p)$!)

Parametric approaches

Fit model to data, read off $I(X;Y)$ numerically (e.g., via Monte Carlo)

Does depend on quality of encoding model, but doesn't depend on Gaussian noise

E.g., instead of discretizing $x \rightarrow x_{discrete}$ and estimating $H(x_{discrete})$, use density estimation technique to estimate density $p(x)$, then read off $H(p(x))$ (Beirlant et al., 1997)

Summary

- Two lower-bound approaches to estimating information
- Very general convergence theorems in discrete case
- Discussion of "bias-correction" techniques
- Some more efficient estimators

References

Beirlant, J., Dudewicz, E., Gyorfi, L., and van der Meulen, E. (1997). Nonparametric entropy estimation: an overview. *International Journal of the Mathematical Statistics Sciences*, 6:17–39.

Miller, G. (1955). Note on the bias of information estimates. In *Information theory in psychology II-B*, pages 95–100.

Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. *Advances in neural information processing*, 14.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253.

Paninski, L. (2004). Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50:2200–2203.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1997). *Spikes: Exploring the neural code*. MIT Press, Cambridge.

Strong, S. Koberle, R., de Ruyter van Steveninck R., and Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters*, 80:197–202.

Treves, A. and Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7:399–407.

Warland, D., Reinagel, P., and Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, 78:2336–2350.

Wolpert, D. and Wolf, D. (1995). Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52:6841–6854.